DOCUMENT RESUME

ED 238 931 TM 840 033

AUTHOR White, Karl; And Others
TITLE An Evaluation of Training in Standardized Achievement Test Taking and Administration. Final Report of the 1981-82 Utah State Refinements to the ESEA Title I Evaluation and Reporting System.
INSTITUTION Utah State Office of Education, Salt Lake City.; Utah State Univ., Logan.
SPONS AGENCY Department of Education, Washington, DC.
PUB DATE Feb 83
CONTRACT 300810271
NOTE 389p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC16 Plus Postage.
DESCRIPTORS Achievement Tests; Elementary Secondary Education; Program Development; Program Evaluation; Program Implementation; *Scores; *Standardized Tests; Student Motivation; Test Coaching; Test Format; *Testing; *Testing Problems; *Test Wiseness
IDENTIFIERS *Confounding Variables; *Title I Evaluation and Reporting System; Utah

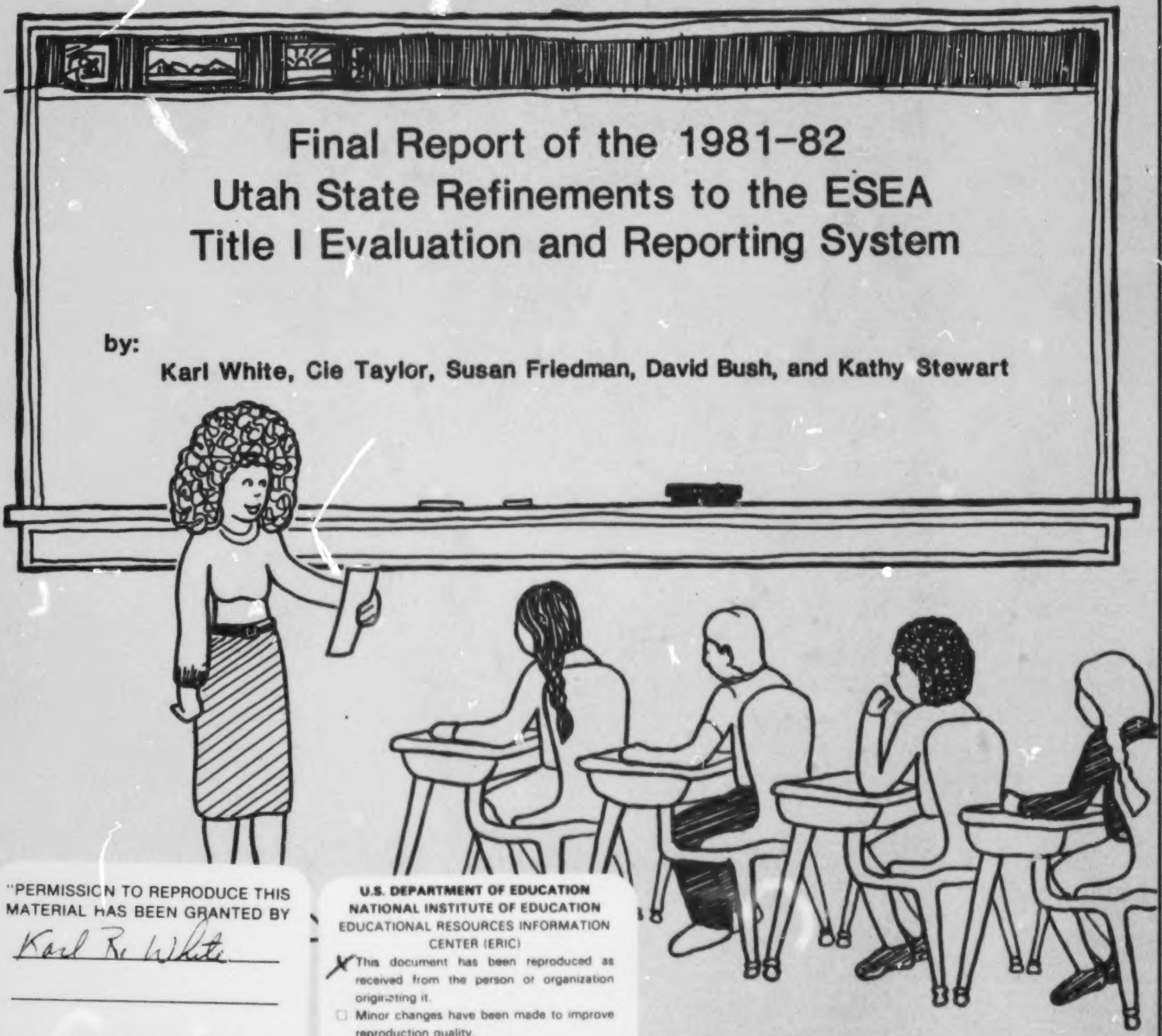
ABSTRACT

Based on findings of previous research, this project developed, implemented, and examined the effect of instructional materials and procedures designed to eliminate the influence on test scores of the following four factors: (1) differential levels of test-taking skills on the part of students; (2) student's lack of familiarity with and consequent confusion from the question format used in the district's standardized achievement tests; (3) lack of motivation on the part of students to do their best on the standardized achievement tests; and (4) inappropriate administration of the standardized achievement tests. Materials and procedures designed to reduce or eliminate the influence of these confounding factors included a series of filmstrips, audiotapes, and workbooks to teach students test-taking skills; a set of seven practice tests; a self-charting procedure designed to motivate students' test-taking; and workshops and exercises to improve teachers' skills as standardized test administrators. Each of these components is described in detail. The contradictions with previous research and teachers' perceptions of the value of the materials and procedures suggest further evaluation of the project materials is necessary before final conclusions are drawn. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED238931

An Evaluation of Training in Standardized Achievement Test Taking and Administration



Final Report of the 1981-82 Utah State Refinements to the ESEA Title I Evaluation and Reporting System

by:

Karl White, Cie Taylor, Susan Friedman, David Bush, and Kathy Stewart

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY
Karl R. White

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Project conducted by:

Utah State University and Utah State Office of Education

TM 840 033

FINAL REPORT

of

STATE REFINEMENTS TO THE ESEA
TITLE I
EVALUATION AND REPORTING SYSTEM :
UTAH 1981-82 PROJECT

FEBRUARY, 1983

ACKNOWLEDGMENTS

The Research and Development work reported in this document would not have been possible without the help of literally hundreds of people. Although all of them cannot be mentioned by name, a few deserve special mention. First, Darrel Allington of Granite School District was the creative drive behind Professor Owl and the unique format of the filmstrips used in the student training materials. His extra efforts in the beginning of the project and patient tutoring and assistance in the later stages is what brought the student training materials into existence. Secondly, the District Coordinators in Granite (Maurice Wilkinson), Nebo (William Rust), Cache (Keith Clayson), and Logan (Gary Carlston) districts who paved the way for their district's participation and served as the liaison person between the project and the district. Third, the dozens of individual teachers who added another concern to an already busy schedule by participating in the project. Their patient understanding and willing participation in spite of changing schedules and unforeseen mishaps was what really enabled the project to be completed. Fourth, the staff at the Utah State Office of Education, particularly Jay Donaldson, Bill Cowan, and Kent Worthington, who provided constant support and encouragement. And finally, the other members of the project staff who do not appear as authors of this final report but did devote hundreds of hours to assuring a quality product, implementation, and evaluation--Marilyn Tinnakul, J. C. Cole, Byron Bair, Edward Konat, and Heather Nairn.

The materials developed in this project and the contents of this report were supported in part by funds from the U.S. Department of Education in conjunction with Contract No. 300810271 (State Refinements to the ESEA Title I Evaluation and Reporting System). The contents do not necessarily reflect the views or policies of the Department of Education, nor does the mention of trade names, commercial products, or organizations reflect endorsement by the U.S. Government.

TABLE OF CONTENTS

Page

EXECUTIVE SUMMARY i

CHAPTER

I. OVERVIEW OF STUDY 1

Objectives 5
Importance and Benefits of Project 6

II. REVIEW OF RELATED RESEARCH 10

Procedures 10

Meta-Analysis Defined 11
Procedures for Meta-Analysis 14
Previous Reviews 17

Meta-Analysis of Research on Reinforcement 19

A Typical Study 20
Results of the "Reinforcement" Meta-Analysis 21
Summary 32
Conclusions 34

Meta-Analysis of Research on Training Students in Test-Taking 37

Definition of Training 38
Previous Reviews 42
Typical Studies 45
Results of the Meta-Analysis 46
Summary 60
Conclusions 61

Review of Research Related to Training Teachers in Test Administration 62

Previous Reviews 63
Test Anxiety 66
Examiner/Examinee Relationships 74
Information About the Test 78
Mechanics of Test Taking 81
Environmental Factors 85
Summary 87
Conclusions 87

Summary 89

TABLE OF CONTENTS (continued)

Page

CHAPTER

III. PROCEDURES 91

Description of Materials 91

Filmstrips and Workbooks: Teaching Students How to Take Tests 92
Practice Tests 112
Reinforcement Procedures 118
Training Teachers in Standardized Test Administration 122
Administration Procedures Specific to a Particular Standardized Test 127
Summary 128

Sample for Research 129

Identification and Selection of Sample 129
Assignment of Sample to Groups 133

Implementation of Experimental Treatments 135

Teacher Training in Test Administration 135
Student Training in Test-Taking Skills 144

Instrumentation 169

Standardized Achievement Tests 169
Student and Teacher On-Task Behavior 173
Locally Developed Instruments 183
Summary 189

IV. RESULTS AND CONCLUSIONS 192

Effectiveness of Training Materials 193

Teachers' Perceptions 193
Differences Between Groups on Outcome Variables 196

Conclusions 217

Implementation Factors Possibly Related to Results 220
Suggestions for Future Research 231

REFERENCES 234

TABLE OF CONTENTS (continued)

Page

APPENDICES

A. Materials Related to Review Literature 245
B. Materials Related to Development of Filmstrips 249
C. Materials Related to Development of Practice Tests 257
D. Materials Related to Reinforcement Procedures 280
E. Materials Related to Sample Selection and Description . . . 283
F. Materials Related to Implementation of Training Materials 288
G. Materials Related to Instrumentation 292
H. Supplementary Data About Effect of Intervention on Dependent Variables 327

LIST OF TABLES

Table Page

1. Summary of Previous Reviews 18

2. Categories for Describing Reinforcement Studies, Number of Effects in Each Category, and Mean Effect Size 22

3. Mean Effect Size by IQ for Contingency, Quality, Age, Design, Test Type, and Number of Subjects 26

4. Mean ES by Unit of Test Administration for Type of Test 28

5. Categories for Describing Student Training Studies, Number of Studies in Each Category, and the Mean Effect Size 49

6. Mean ES by Quality for Type of Training and Number of Subjects 51

7. Mean ES by Type of Test and Quality of Research Design for Type of Training, Unit, Age, Design, and IQ 53

8. Mean ES for Studies Coded "Achievement" by Quality, Test, Age, and Training 54

9. Use of Tests in Utah Title I Projects 93

10. Number of Test Adoptions for Districts and States by Region 94

11. Summary of Test Use for Project Tests 95

12. Subtests 96

13. Objectives for Student Training Filmstrips 97

14. Time Line for Making Filmstrips and Tapes 111

15. The Mean Number of Items and Minutes Used for Each Practice Test 113

16. Computations Used to Determine Number of Items to Include in a 30-Minute Practice Test 115

17. Description of Districts Participating in Project 130

18. Experimental Sample 131

19. Implementation of Experimental Treatments 135

20. Actual Timeline for Implementing Filmstrips, Practice Tests, and Teacher Supervision 136

LIST OF TABLES (continued)

Table ... Page

21. Training in Test Administration Breakdown by District ... 137

22. Fall Workshop Evaluation Data ... 141

23. Workshop in Student Training Implementation Breakdown by District ... 145

24. Teacher Training in Student Curriculum Workshop ... 147

25. Number of Contacts Between Project Staff and District Staff ... 153

26. Number of Classrooms, Students Present, and Students Absent for Each Filmstrip ... 155

27. Number of Classrooms, Students Present, and Students Absent for Each Practice Test ... 156

28. Summary of Filmstrip Evaluations ... 158

29. Results from Teacher Evaluation: Project Components ... 161

30. Verbal Comments from Teachers on Project ... 164

31. Mean Ratings Given Teachers for Support and Quality ... 168

32. Standardized Test Formats ... 171

33. Test Statistics on Data Collection Instruments Developed by Project ... 176

34. Breakdown for Observations by Number of Classes ... 181

35. Practice Data Collection: Percent of Interrater Agreement for Quality of Test Administration ... 182

36. Practice Data Collection: Percent of Interrater Agreement for On-Task Behavior During Teacher and Student Directed Tests ... 182

37. Actual Data Collection: Percent of Interrater Agreement for Quality of Test Administration ... 184

38. Actual Data Collection: Percent of Interrater Agreement for On-Task Behavior During Teacher and Student Directed Tests ... 184

39. Intercorrelations of Dependent Measures ... 191

LIST OF TABLES (continued)

Table Page

40. Scores on Dependent Variables by Experimental Group 198

41. Total Achievement Test Scores by Group by Various Independent Measures 214

42. Intercorrelation Matrix for Project Variables 218

43. Third Grade Standardized Achievement Test Scores from 1981-82 Year 224

LIST OF FIGURES

Figure Page

1. Distribution of 41 Effect Sizes for Reinforcement Studies Considered in the Meta-Analysis 24

2. Distribution of 62 Effect Sizes from Student Training Studies Considered in the Meta-Analysis 48

3. The Time Period for Producing Each Practice Test 113

4. Basic Definitions for On-Task Behavior 178

5. Box and Whisker Diagrams and Normal Curve Representations for Major Dependent Variables 208

6. Box and Whisker Diagrams for Dependent Variables Using Teacher as Unit of Analysis 226

EXECUTIVE SUMMARY

Testing is a major industry for the American educational system. The National Education Association estimates that approximately 200 million achievement tests are administered annually in the United States (McKenna, 1973). Of the three or four published tests students take each year, the majority are standardized measures. Scores from these tests are used as a primary source of information in making decisions about educational programming, class placement, student advancement, and evaluations of educational programs. Given the fact that standardized achievement tests are used to make such important decisions, it is essential to be sure the tests are really measuring what they purport to measure.

Unfortunately, previous research suggests that several other factors may be confounded with scores received by students on standardized achievement tests. To the degree that such factors are influencing the scores students receive, decisions which are based on the results of standardized achievement tests may be misleading and/or inappropriate. Some of the potentially confounding factors identified in previous research include the following:

- Administration Procedures. Teachers who do not follow standardized test administration procedures in administering the test may cause students' scores to be higher or lower than they would otherwise be. Students may receive higher scores than they deserve if the test is not properly monitored, if inappropriate hints or assistance are given, or if cheating is not carefully controlled. Students may receive lower scores than they should if directions are not given clearly, if they are not properly prepared for test, or if a nonsupportive or anxiety-provoking atmosphere is maintained.
- Student Test-Taking Skills. Several previous research studies have suggested that mastery of test-taking skills such as checking work, using elimination strategies, timing, and following directions are positively related to student scores.
- Test Format. The format used by different standardized achievement tests to assess a student's mastery of the same content area is often radically different. There are indications that unfamiliar test

formats may be confusing for students. Such confusion may result in lower test scores even though students know the material being tested.

- Student Motivation. Students who are not motivated to do well on standardized achievement tests will probably receive lower scores than they would have if they had tried their best. When this happens, the test is at least partly a measure of student motivation, even though the decisions based on test scores assume that the test is solely a measure of achievement or mastery of a particular content area.

Objectives

To the degree that these previous research findings are correct, student scores on standardized achievement tests will be invalid because factors other than what the student knows (e.g., familiarity with format, administration procedures, motivation, and test-taking skills) will influence scores on the test. Based on the findings of previous research described above, this project developed, implemented, and examined the effect of instructional materials and procedures designed to eliminate the influence on test scores of the following four factors.

1. Differential levels of test-taking skills on the part of students.
2. Student's lack of familiarity with and consequent confusion from the question format used in the district's standardized achievement tests.
3. Lack of motivation on the part of students to do their best on the standardized achievement tests.
4. Inappropriate administration of the standardized achievement tests.

Materials and Procedures

Materials and procedures designed to reduce or eliminate the influence of the confounding factors described above included a series of nine filmstrips, audiotapes, and workbooks to teach students test-taking skills; a set of seven practice tests formatted similarly to the standardized achievement test used by the district; a self-charting procedure designed to motivate students to try their best on tests; and, workshops and exercises designed to improve teachers' skills as standardized test administrators. Each of these components is summarized briefly below and described in detail in the project's Final Technical Report.[1]

Filmstrips for teaching test-taking skills. Nine filmstrips (lasting approximately 30 minutes each), audiotapes, and workbooks were developed to teach students test-taking skills such as checking their work, filling in answer spaces correctly, following directions, differentiating between correct and look-a-like answers, using different question formats, and using partial knowledge to eliminate wrong answers. All of the test-taking skills instruction focused on standardized reading achievement tests. Filmstrips were designed so that the lights remained on during the filmstrip and the following instructional principles were emphasized:

- the teacher interacted with the filmstrip, controlling the pace of instruction, checking student mastery, supplementing instruction when necessary, and demonstrating correct performance.
- students were actively involved in the instruction--completion of the workbooks occurred during the filmstrip, vocal responses were used frequently, and teachers were instructed not to proceed to new material until all students had demonstrated mastery.

Practice tests. Seven practice tests (ranging in length from 5 to 30 minutes) were developed using the same format used by each district's standardized achievement test. Content of the practice tests paralleled what was being taught to students in their reading group during the year. The practice tests provided an opportunity for students to practice the concepts being taught in the filmstrips, become familiar with the format and standardized testing procedures (teachers administered each practice test using written directions similar to a standardized test), and to learn to work independently in a standardized testing situation. By the time students took the standardized test in the spring, it was hoped that standardized testing procedures would be a familiar and comfortable experience. In addition, these practice tests were designed so teachers obtained feedback at periodic intervals on student mastery which could be useful in designing their classroom instruction.

---

[1] Copies of this report are available from the United States Department of Education (Reference Contract #300810271), the Utah State Office of Education, or the ERIC Document Reproduction Service.

Motivating students to try their best. Scores on the practice tests were also used as a basis to motivate students to try their best on standardized achievement tests. A self-charting procedure was used with each student's individual chart prominently displayed. Each student received points to be put on the chart for improving his or her score from the previous practice test (students scoring above 80% on each practice test were always given points). It was hypothesized that if students learned to try their best on the "standardized" practice tests, this motivation would transfer to the actual standardized achievement test given in the spring.

Training teachers in standardized test administration. Teachers were trained in two workshops (one in the fall, one in the spring) to be better standardized test administrators. During the workshops, teachers were instructed in standardized testing procedures, critiqued videotapes of good and bad test administration, and role played various aspects of test administration. Examples of the type of concepts emphasized included student seating arrangements, preparing for early finishers, clarifying ambiguous directions and making sure all students understand directions, and facilitating a supportive and properly controlled atmosphere.

Experimental Design

To test the effect of the training materials on students' and teachers' performance during standardized achievement tests, 58 classrooms from three school districts in Utah were randomly assigned to one of three groups.

- Experimental Group 1 classrooms received all of the training materials (filmstrips, practice tests, motivation procedures and training in test administration).
- Experimental Group 2 classrooms received only the student training materials (filmstrips and practice tests).
- Control group classrooms received no specially prepared materials concerning the administration of the standardized achievement tests.

Project staff at Utah State University provided extensive supervision and assistance to each of the experimental group classrooms participating in the project including training workshops, on-site modeling of material, periodic on-site follow-up and assistance, and telephone consultation. Each teacher in the experimental groups was visited an average of five times during the year in addition to the training workshops. Also, there was an average of 7.9 phone consultations with each of the teachers.

The effectiveness of the project in teaching elementary school students test-taking skills, motivating students to do their best on standardized achievement tests, and training teachers in the proper administration of standardized achievement tests was assessed based on data collected for each of the three groups in the following areas:

1. Teachers' responses to questionnaires and interviews to assess their perceptions of the value of the materials and the quality of implementation.
2. Students' scores on the district's standardized achievement test.
3. Observations by blind observers of student and teacher on-task behavior during the standardized achievement test.
4. Student and teacher attitudes towards standardized achievement tests as measured by paper-and-pencil attitude instruments developed by the project.

5. Ratings by blind observers of the quality of the teacher's test administration.

The actual measures used to collect data about the project are described in greater detail in the Final Technical Report of the project. These measures consisted of a series of standardized and locally developed measures and observation systems. The standardized achievement tests were administered in each class by the classroom teacher as was the general practice of the participating districts. All other data were collected by specifically trained data collectors who were uninformed as to the nature of the research or the group membership of any other classes.

Results

Teachers' perceptions. Most components of the project, particularly filmstrips and practice tests, were viewed very positively by teachers. For example:

- 84.2% of the teachers felt the filmstrips were worth the time and effort required.
- 78.9% plan to use the filmstrips next year.
- 94.7% felt the filmstrips taught concepts which were important for students to learn.
- 79% felt the practice tests adequately prepared the students for taking the standardized achievement test.
- 76.3% plan to use the practice tests in the future.
- 76.3% felt the benefits of the total project were worth the investment in time.
- 73.7% of the teachers felt the project was enjoyable for students.
- 81.6% of the teachers felt the project benefited the student's test-taking skills.

Teachers' perceptions of the value of the procedures for teaching standardized test administration skills were also very positive. Seventy-one percent of the participating teachers felt they were better test administrators as a result of the workshops. However, the procedures used to motivate students to try their best on tests were viewed less positively. Slightly more than half the teachers (53.3%) felt that the motivational procedures were difficult for students to understand. Only about a third (38.1%) felt that the students were motivated by the procedures, and only 38.3% of the teachers plan to use the motivational procedures in the future.

Teacher and student attitudes and behaviors. Table 1 includes data for all of the major outcome measures for each of the three groups ("1" refers to Experimental Group 1, "2" refers to Experimental Group 2, and "C" refers to the control group). As can be noted, teachers participating in the project had improved attitudes towards standardized achievement tests, particularly those teachers in Experimental Group 1 who received the training in standardized test administration. Teachers' on-task behavior and quality of test administration was also significantly improved as a result of the project. Differences between groups on student attitudes towards standardized tests were statistically significant favoring the control group. However, in practical terms, these differences were very small. There were no differences between the groups on student on-task behavior during the test.

Academic achievement. There were statistically significant differences between the groups on all of the achievement test scores with Experimental Group 2 scoring the highest. Although statistically significant, differences between the groups are relatively small (an average of less than one-quarter a standard deviation). However, it is important to note that students in Experimental Group 1 who received the most intervention consistently scored the lowest on the standardized achievement test scores.

Table 1

Scores for Each Group on Major Dependent Measures

Variable	Scores and Rank Order						p^b	Cont. Grp. SD	ES^c
Teacher Attitude	$\bar{X}$	85.7		87.9		89.1	.000	12.3	.35
		C	<	2	<	1			
	Md	85.5		87.2		89.8			
Student Attitude	$\bar{X}$	11.7		11.9		12.4	.011	3.5	.20
		2	<	1	<	C			
	Md	11.2		11.3		12.1			
Teacher On-Task (Teacher-Directed Subtest)[a]	$\bar{X}$	59.2		73.1		77.1	.000	49.0	.60
		C	<	2	<	1			
	Md	68.9		89.1		98.3			
Teacher On-Task (Student-Directed Subtest)[a]	$\bar{X}$	83.2		78.4		80.6	.048	37.7	.14
		C	<	2	<	1			
	Md	87.7		88.7		92.8			
Student On-Task (Teacher-Directed Subtest)	$\bar{X}$	88.4		89.2		89.7	.785	11.1	.32
		1	<	C	<	2			
	Md	90.8		92.5		94.4			
Student On-Task (Student-Directed Subtest)	$\bar{X}$	90.5		90.6		89.9	.911	9.3	.14
		C	<	1	<	2			
	Md	92.5		93.4		93.8			
Quality of Test Administration Rating	$\bar{X}$	48.8		50.6		49.7	.000	3.8	.28
		C	<	2	<	1			
	Md	48.3		50.9		52.1			
Achievement Test (Teacher-Directed Reading Subtest)	$\bar{X}$	-.10		.04		.07	.023	.9	.24
		1	<	C	<	2			
	Md	.06		.09		.32			
Achievement Test (Student-Directed Reading Subtest)	$\bar{X}$	-.08		.08		.01	.033	.9	.19
		1	<	C	<	2			
	Md	.19		.33		.38			
Achievement Test (Total Reading)	$\bar{X}$	-.10		.07		.05	.013	.9	.29
		1	<	C	<	2			
	Md	.05		.22		.34			

[a] For each standardized achievement test, a teacher-directed subtest was defined as one where the teacher gave directions and controlled the pace item by item. A student-directed subtest was defined as one where directions were given at the beginning of the subtest and then student worked independently for a certain time limit or until they finished.

[b] All probability estimates are based on one-way analyses of variance between means of the three groups. In many cases, distributions are substantially skewed so that medians are a better indicator of central tendency. Medians for each group on all variables are also reported. Asterisks are used to indicate where the order of groups differs depending on whether means or medians are reported. The order of groups represented in the chart always follows medians.

[c] The column labeled ES refers to the standardized mean differences between the highest and lowest group $(\bar{X}_{high} - \bar{X}_{low}) \div SD_{control\ group}$. This measure has been recommended by Glass (1977) for examining the results of various studies using a common metric.

Subgroup analyses. To check the robustness of the findings reported above when all students were included in the analyses, several additional analyses were done using subgroups of students. These subgroup analyses were done for the following groups of students across all dependent variables.

- Students who received the majority of the experimental treatment. These analyses were done eliminating those students who saw less than 5 filmstrips, took less than 3 practice tests, had teachers who were rated low on quality of implementation or support, were in special education programs, or had English as a second language.
- Students who received all of the experimental treatment. These analyses were done eliminating those students who saw less than 9 filmstrips, took less than 7 practice tests, had teachers who were rated low on quality of implementation or support, were in special education programs, or had English as a second language.
- Only Title I students who received all of the treatment.
- Students in each of the three participating districts analyzed separately.

The results of the subanalyses (reported in detail in the project's Final Technical Report) confirmed in all cases the results reported above in Table 1.

Conclusions

The purpose of this project was to develop, implement, and evaluate the effect of training materials and procedures designed to increase the validity of standardized achievement tests by improving:

- Students' test-taking skills, attitudes toward tests, and motivation, and
- teachers' attitudes toward standardized tests and quality of test administration.

As noted briefly above, the intervention procedures did result in improved teachers' attitudes towards tests and quality of test administration. Furthermore, teachers were enthusiastically supportive of the materials, plan to continue using the materials in the future, and felt that the materials resulted in substantial improvements in students' test-taking abilities and students' attitudes towards tests. However, the more objective data collected by the project indicated that there were no meaningful increases in students' test-taking skills or students' attitude or performance during tests.

These data raise some perplexing questions in view of previous research which has supported the efficacy of the types of intervention developed in this project, and in view of teachers' perceptions about the effectiveness of the project. First, previous research has indicated that training students in test-taking skills has a substantial effect on test scores. When compared to the interventions in previous research, the training delivered to students in this project was a relatively intense, systematically delivered training experience of long duration with good follow-up and monitoring. In spite of this, no meaningful differences were observed between the groups on test scores. Most differences which were observed were not in the predicted direction. In fact, those students who received the most training received the lowest scores.

The fact that differences were not found is even more perplexing in light of teachers' very positive response to the program materials. Most teachers who used the materials during this year plan to continue using the materials in the future and felt that the materials had improved their students' attitude and increased performance on standardized achievement tests. However, the fact remains that none of these perceived differences were apparent on objective measures for which data were collected on the project.

The contradictions with previous research and teachers' perceptions of the value of the materials and procedures suggest that further evaluation of the materials developed in this project should be conducted before final conclusions are drawn. Further research is necessary to understand to what degree typically administered standardized achievement tests are valid and useful for the purposes for which they are usually used. The materials developed in this project represent an important beginning. As they are used further and more data are collected, we will be able to better understand the degree to which results from standardized tests should and can be used to make programming, evaluation, and placement decisions for primary grade children.

FINAL REPORT:

STATE REFINEMENTS TO THE ESEA, TITLE I EVALUATION AND REPORTING SYSTEM

CHAPTER I

1. OVERVIEW OF THE STUDY

The general purpose of RFP 81-034 was to support further development work by State Education Agencies SEAs to "enhance the quality of Title I evaluation at the state and school district level." The project described in this report accomplished this overall objective by addressing the following two specific areas targeted by RFP 81-034.

- Quality Control. Efforts designed to improve the accuracy and validity of the Title I evaluation data currently being collected.
- Measurement and Evaluation. Studies, designed to investigate technical aspects of the current evaluation models . . .

The Title I Evaluation and Reporting System (TIERS) was designed to provide decision makers at all levels with information about:

> ". . . the effectiveness of the programs assisted under this Title in meeting the special educational needs of educationally deprived children; . . . such evaluations will include . . . objective measurement of educational achievement in basic skills . . ." (ESEA, Title I, Section 124 (G)).[1]

In other words, TIERS was designed to provide information about how much more children know in basic skills areas than they would have known had they not participated in Title I programs. Each of the TIERS Models utilizes standardized achievement tests to provide information about how much children know about basic skills. Each of the models compare children's scores on the achievement tests at the end of the program with a no-treatment expectation (i.e., what children would have known had they not participated in the program). The difference in these two estimates of children's knowledge is assumed to be attributable to the effect of the Title I program.

---

[1] For a more complete description of TIERS, see Tallmadge and Wood (1981).

In addition to using scores from standardized achievement tests to evaluate the impact of Title I programs, most Title I projects also use standardized achievement test scores in selecting children to participate in Title I programs, and in making educational programming decisions about those students once they have been placed in the program.

The validity of these decisions (i.e., decisions about program impact, student placement, and programming for students) depends on the scores from the standardized achievement tests actually measuring what the user of the test results thinks it is measuring (which in most cases is the student's knowledge of the basic skill area being tested). In other words, for the results of TIERS to be useful, valid standardized test results must be obtained. However, Cassell (1969) noted that there are at least two conditions critical to obtaining valid standardized test results:

1) The student's score on the test must be a function of what the student knows about a topic rather than some other variable.
2) The test must be administered according to specified standardized procedures.

There are difficulties associated with the failure to meet either of these conditions.

An example of violating the first condition occurs when a test is a valid instrument for one purpose or in one setting, but does not yield a valid score for the particular setting or purpose for which it is being used. Variables such as a student's test-wiseness or test-taking skills and level of motivation may influence a test score so that an accurate estimate of academic achievement for that particular student cannot be obtained. For instance, if a student fills in the bubble on the machine-scorable form too lightly to be read, the resulting score will be lower than if the machine had read all of the answers; or, if a student doesn't feel like taking a test that day, the score will be different than on a day on which the student is motivated to do his/her best.

The validity of standardized test scores are also called into question when a test administrator violates standardized administration procedures (the second condition referred to above). When standardized test administration procedures are not followed, children may misunderstand the directions, cheating may occur, or time limits may be altered. Additionally, when standardized administration procedures are not followed, the comparison of obtained scores to those of the norming group will probably be inappropriate. For example, test results obtained by students who did not receive a practice test prior to the actual test may provide diagnostic information, but will not be interpretable according to a norm group if the students in the norm group did take a practice test.

Despite the importance of these factors in obtaining valid and interpretable scores, it appears that little is being done in many classrooms to assure that:

1) tests are indeed measuring academic skills (and not level of test taking skills or motivation); and,
2) standardized test administration procedures are followed.

Most test companies encourage the use of standardized procedures by including a section in the test manuals to alert teachers of the importance of following standardized directions. Furthermore, most teachers have received some training in the administration of standardized tests, and most people recognize the importance of selecting tests which are appropriately matched with the school's instructional emphasis and encouraging students to do their best. However, data collected by the Utah State Office of Education during the 1979-1980 school year as one part of the previous project for State Refinements to the ESEA Title I Evaluation and Reporting System (White, Taylor, Eldred, & Carcelli, 1981; hereafter referred to as "79-80 State Refinements Project"), indicated that substantial problems exist in Utah Title I programs which make the interpretation of Title I evaluations, regardless of which one of the TIERS models is used, difficult and perhaps misleading.

The 79-80 State Refinements Project identified four primary factors which may be confounding the results of Title I evaluations designed to estimate how much more students know as a result of Title I programs than they would have known had they not participated in the program. These factors included:

1) the procedures used during test administration;
2) the test-taking skills of the students;
3) the format of the particular standardized achievement test which is used; and
4) the motivational level of the students.

Data from the 79-80 State Refinements Project provided evidence that existing problems in each of these four areas may confound the interpretation of Title I evaluation results. The project described in this report expanded on the previous project to (a) more definitively investigate the causal relationship of the above factors with student test scores; and, (b) design, implement, and test the effectiveness of procedures designed to reduce or eliminate factors in each of these areas which may confound the interpretation of scores from standardized achievement tests. The project was a cooperative effort between the Utah State Office of Education, researchers at Utah State University, and four LEAs within the State of Utah. The remainder of this section outlines the specific objectives of the project and explains the importance and potential benefits of the study.

Objectives

The overall goal of this study was to design, implement, and test the effectiveness of procedures and training packages designed to increase the validity of standardized achievement test scores typically used throughout Utah in implementing the Title I Evaluation and Reporting System. The projected benefit of such development and evaluation work was the increased validity of results from the Title I Evaluation and Reporting System as a consequence of standardized testing procedures being followed more rigorously and confounding factors such as test-wiseness, motivation of students, and testing format being reduced or eliminated. The specific objectives of the study included:

1) LEA personnel administering standardized achievement tests used in Title I Evaluation and Reporting System will adhere more closely to standardized testing procedures and will display more positive attitudes and increased skill consistent with standardized testing procedures in administering the tests.
2) Students will be more motivated to take standardized achievement tests and will display higher levels of test-taking skills which will eliminate these factors as confounding variables in demonstrating what students know.
3) The confounding effects on student test scores of question format will be eliminated or reduced.
4) The causal relationship between scores on standardized achievement tests and quality of test administration, student test-taking skills, student motivation, and item format will be determined.

These objectives were addressed by designing, implementing, and evaluating the effectiveness of experimental treatments for students in Title I schools. Experimental treatments consisted of:

1) Training teachers in proper standardized test administration procedures.
2) Training students in test-taking skills.
3) Implementing procedures for motivating students during the regular school year to achieve well on tests.
4) Familiarizing students with the test formats used by their district's standardized test.

Classrooms in Title I schools were randomly assigned to various experimental and control conditions to test the effectiveness of the intervention procedures and to establish what, if any, causal relationship existed between these factors and students' test scores. The effects of the various interventions were investigated using a variety of dependent variables including observation of teacher and student on-task behavior during testing, scores on the Quality of Test Administration Checklist, student and teacher attitudes toward testing, and student scores on the standardized achievement test.

Importance and Benefits of Project

Results of the Title I Evaluation and Reporting System (TIERS) will be invalid or misleading to the extent which factors such as administration procedures, students' test-taking skills, student motivation, and test format are confounding students' scores on standardized achievement tests. As a result of this project which developed and evaluated the effectiveness of the procedures to eliminate or control these variables in Title I evaluation situations, local, state, and national Title I officials can better understand how TIERS results should be interpreted.

As a result of the project, several training packages are available to LEAs to train teachers in the proper administration of standardized achievement tests and train students in test-taking skills. The following materials have been produced and are available from the Department of Education or the Utah State Office of Education for the cost of reproduction.

1) Training Teachers in Test Administration Procedures: Presenter's Guide (150 pp.).
2) Taking Tests: A Little Magic Always Helps (a series of nine filmstrips, work booklets, and audiotapes).
3) How to Take Tests: Teacher's Manual (312 pp., includes masters for workbooks, practice tests, reinforcement procedures, and filmstrip scripts for filmstrip series).

The potential benefits of a project such as this for the U.S. Department of Education are more far reaching. Testing is a major industry in the American educational system. The National Education Association estimates that approximately 200 million achievement tests are administered annually in the U.S. (McKenna, 1973). Of the three or four published tests that students take each year, the majority are standardized measures. Scores from these tests are a major source of data that are used to report low achievement and inequities in the delivery of educational services. If test scores are to be used to document the occurrence of educational inequity, to compare results across groups of students, and to make educational decisions, the test results need to be valid and interpretable as indicators of student knowledge (i.e., scores must be a measure of the skills the test was designed to measure).

Currently, test results are used at every level of education from teaching to formulating policy. The objectives addressed by this project are particularly relevant for four different areas which include but extend well beyond the concerns of Title I: (a) the use of norm referenced evaluation procedures, (b) the placement of students into special programs and curriculum, (c) the diagnosis of academic deficiencies, and (d) the funding and policy making for selected educational groups.

First, the success of instructional programs is often determined by comparing the pre- and posttest scores of the treatment group to scores of an empirically established norming group. Group test scores found to be sensitive to variations in testing procedures and in student motivational levels may not be interpretable according to published normed tables. For example, if a spring pretest and a spring posttest are used to evaluate a program, the two tests were probably given by different teachers and the gain or loss may be attributable as much to the way the two tests were administered as to the effects of the instructional program.

Second, students most affected by the use of group achievement test scores for diagnostic and placement purposes are frequently those with relatively low academic achievement levels or socioeconomic status and are in programs such as special education, bilingual education, or Title I. Many of the remedial groups in which a student may be placed have a limited number of spaces that are filled by students with the greatest need. If the basis for placement in special instructional programs is a low test score, and if the scores of some students are influenced by test-taking skills, low motivation, or improper test administration, selection decisions may be incorrect.

A third area affected by variation in testing conditions is academic assessment. Once students are placed into a program, academic deficiencies should be precisely identified so that valuable instructional time is not spent teaching skills that have already been mastered. If a student's score is a function of misunderstanding of directions, low motivation, or poor test-taking skills, deficiencies will be improperly noted and development may be retarded by incorrect instructional grouping or programming.

Finally, additional knowledge about the factors examined in this study is important for funding and policy decisions that rely on student test scores.

For instance, if the correlation between ethnicity and test scores becomes significantly lower when differences in testing conditions, test-taking skills, and motivational levels are controlled, then ethnic group comparisons made under uncontrolled conditions are less believable. That some data may not be a valid estimate of achievement is particularly disconcerting when the people who use the actual test scores for financial allocations (e.g., legislators) are removed from the test setting and are forced to rely on the "facts" from score reports.

CHAPTER II

REVIEW OF RELATED RESEARCH

Previous authors have suggested that student scores on educational tests may vary as a result of factors other than knowledge of the content being tested and random error (Ebel & Damrin, 1960; Thorndike, 1949; Vernon, 1962). The purpose of this section is to review and synthesize the findings from previous research which were most relevant for the materials and procedures of this project. The discussion of the effect of the following three factors on test score differences among students establishes the theoretical and empirical basis for much of the work described in the Procedures Section.

1) Reinforcement (RE)--giving students verbal or tangible rewards contingent or noncontingent on test scores;
2) Student training in test-taking skills (ST)--providing students with practice, coaching, or training in test-wiseness; and
3) Teacher training in test administration (TT)--training examiners on how to implement standardized procedures and how to prepare students for a test.

The procedures used for conducting the reviews are described first and then relevant studies are grouped and reviewed separately for each of the three factors. A summary of the three reviews is the final section of this chapter.[2]

Procedures

Two approaches were used to review and summarize the results from previous research. First, a "meta-analysis" was conducted on studies of reinforcement and student training. A description and rationale of meta-analysis are presented below. Second, because insufficient research was located on the effects of teacher training to justify using meta-analysis, findings from studies related to the administration of standardized tests are presented. Since this research covers a variety of testing conditions, each study is briefly described and summarized.

---

[2] Additional work by the authors in developing prototypes of some of the materials used in this project is described in the Technical Proposal for RFP 80-034 and in the Final Report of the previous Utah State Refinements contract. The theoretical review in this section does not refer to this work.

The remainder of this section defines meta-analysis, describes the meta-analysis procedures, and discusses previously completed reviews on Reinforcement (RE), Student Training (ST), and Teacher Training (TT).

Meta-Analysis Defined

The term meta-analysis was introduced by Gene Glass in 1976 to describe the statistical analyses performed on the results of individual studies for the purpose of integrating findings. McGaw and White (1981) quote Glass:

> The approach to research integration referred to as "meta-analysis" is nothing more than an attitude of data analysis applied to quantitative summaries of individual experiments. By recording the properties of studies and their findings in quantitative terms, the meta-analysis of research invites one who would integrate numerous and diverse findings to apply the full power of statistical methods to the task. Thus, it is not a technique; rather, it is a perspective that uses many techniques of measurement and statistical analysis. (p. 12)

Hence, meta-analysis is not a new methodology--it is an approach which uses different existing research technologies depending on analyses to be completed. Three characteristics distinguish meta-analysis:

1. The outcomes of individual studies are quantified on a common metric so that results can be compared across studies. Examples of quantification are the use of standardized scores (such as IQ, r^2, omega (ω) squared, eta (η) squared, or Glass' ES, see below).

2. A comprehensive list, or at least, a representative sample of studies is considered (including journals, government reports, dissertations, and unpublished material) so that results of the review can be generalized to research which has been conducted on that topic.
3. Characteristics of individual studies (e.g., size of sample, type of design, and age of students) are quantified and coded so that the covariation of study characteristics and the outcomes can be systematically and empirically examined.

To conduct a meta-analysis, a comprehensive list or a representative sample of studies is identified by clearly specified procedures. Next, the features of the studies are coded quantitatively and outcomes are converted into a common metric. Finally, findings are described and analyzed by statistical procedures to examine the covariation of study characteristics with the outcome measures.

A meta-analysis offers several advantages (see below) over more typical review of research that often present studies with differing results and no conclusions. Glass (1977) summarized the problems frequently encountered by review of existing research:

1. Literature searches are haphazard and selective, and often omit dissertations.
2. Reviews are typically narrative and discursive; findings are often difficult to understand without quantification.
3. Reviewers who attempt to quantify studies generally (and inappropriately) use statistical significance as a method of integrating studies to draw conclusions.

Meta-analysis, though not a panacea for all review ills, does solve many difficulties associated with traditional reviews by being somewhat more unprejudiced, quantitative, and generalizable. First, it is unprejudiced because studies are not arbitrarily or selectively excluded on the basis of quality (e.g., poor design, questionable implementation, inappropriate dependent variables). Instead, a representative sample of previously completed research is considered and characteristics of design and analyses which contribute to the quality (i.e. good vs. bad) of the research are simply coded for use in further analysis.

Second, meta-analysis is quantitative because outcomes from large numbers of studies can be organized by using the same metric. This common metric, referred to as effect size (ES), is usually defined as

$$\text{Effect Size} = \frac{\bar{X}_{\text{experimental}} - \bar{X}_{\text{control}}}{SD_{\text{control}}} \quad \text{(Glass, 1977).} \qquad (1)$$

Finally, meta-analysis yields more generalizable results because the studies selected for use in the meta analysis must be comprehensive (include all the research) or be representative (randomly sampled to typify all research). In addition, the relevant characteristics of each study are coded and entered into the analysis as variables. This process encourages stronger, more adequately supported conclusions than reviews which synthesize research on the basis of methodology or statistical significance.

Procedures for Meta-Analysis

This section describes the procedures used to locate studies and code study characteristics for the meta-analysis. The steps explained below are those used to complete two separate meta-analysis, one analysis for reinforcement and another analysis for student training. Specific details about the individual analyses will be provided in the sections Reinforcement and Student Training.

Locating studies. The first step was to collect all the studies regarding reinforcing test behaviors (RE), student training in test-taking skills (ST), or teacher training in test administration (TT). The primary sources for these studies were four library data based computer searches conducted at Utah State University. The data bases included Educational Resources Information Center (ERIC), Current Index to Journals in Education (CIJE), Psychological Abstracts, and Dissertation Abstracts. Computer searches yielded 31 RE, 79 ST, and 0 TT titles by using combinations of the following descriptors.

test (ing) (s)
administration (tor)
reinforce (r) (ment)
train (ing)
standardize (d)
intelligence (IQ)
test wiseness (TW)
elementary
test score (s)
motivation
reward (s)
achievement
student
teacher
exam (iner) (ation)
practice
coaching
aptitude

Since no research was located for the TT factor, this review was dropped from the meta-analysis. Once the articles for RE and ST were located, the bibliographies and references provided a second source of studies.

To qualify for inclusion in the meta-analysis, articles had to meet the following criteria:

1. The test used to measure the outcome had to be a standardized intelligence or achievement test (aptitude tests were classified as achievement).

2. At least one independent variable had to be a "treatment" applied to subjects.

3. The outcome data had to be reported as test scores.

4. The research could not be supported by a test publisher because of the possible bias that might ensue during the study.

Some of the articles failed to meet criteria and were therefore excluded. The most common deficiencies found in rejected studies were the use of a nonstandardized outcome measure (n = 17) and researcher affiliation with a test publisher (n = 27).

Some articles described more than one treatment effect and each effect within an article was separately coded. For example, if the impact of practice testing was measured twice (immediately following the practice and after one month), scores from both posttests were used to compute two effects. The final yield was 41 RE effects from 18 articles and 62 ST effects from 37 articles. The 55 articles used in the meta-analyses are identified in the References as "R" for reinforcement and "T" for student training.

Coding study characteristics. The next step in the meta-analyses was to describe the study characteristics. To determine what information to include in the analyses, all the studies were read and preliminary estimates were made as to which research conditions might affect the relationship between test scores and reinforcement or student training. These conditions were listed on a summary sheet so that each article could be quantified on various characteristics. Examples of the coding sheets used for RE and ST are located in Appendix A. The following study characteristics were coded for both Reinforcement studies and Student Training studies: number of subjects, mean age of subjects, mean IQ of subjects, type of tests used as a dependent variable (IQ, achievement, or aptitude), test administration unit (group or individual), type of research design, quality of research design, effect size, and investigator's conclusion about the effectiveness of the intervention. For the Reinforcement studies, the type of reinforcement (money, candy, praise, reproof, token, choice, and prize), the schedule (immediate or delayed), and the contingency (contingent or noncontingent on correct test scores) were also coded. The type of training (practice or test wiseness) provided was coded for the Student Training studies.

"High" quality was coded when studies basically accounted for internal and external threats to validity (Bracht & Glass, 1968; Campbell & Stanley, 1963). "Low" quality was assigned to studies that failed to control for one or more major extraneous variables.

One or more effect sizes (ES) were computed using Equation 1, where each mean student test score was transformed for each study into a common index that described the impact of the intervention (reinforcement or student training) and could be compared across studies. For studies that did not report standard deviations or means on the outcome variable, the

ES was calculated from statistics such as t or F, using procedures outlined by McGaw and Glass (1981). For pre-post studies where one group was compared against itself, the pre-test mean score was used in the calculations as the control group mean.

Previous Reviews

Seven previous reviews on test-taking skills or test administration were located and the characteristics of each review are summarized in Table 1. Five reviews summarized research on training students in test wiseness. Two reviews described studies which examined procedures related to test administration even though no studies examined the effects of training examiners. No reviews were located on reinforcing testing behaviors.

In general, the previous reviews lacked three critical components (Glass & Smith, 1979): (a) a systematic method for identifying studies to be included; (b) a common index for quantifying data for comparisons across studies; and (c) a systematic integration of the reviewed data into a meaningful summary.

Although one reviewer (Vernon, 1954) reported results in effect size by converting all mean scores to IQ units, the covariation of study characteristics with outcomes was not considered systematically. Sattler and Theve (1967) described research in terms of level of significance and drew conclusions by "voting" (see Light & Smith, 1971) on the number of studies that obtained statistical significance versus the number of studies with nonsignificance. The remaining five reviews discussed studies in terms of conclusions drawn by the primary researcher.

Table 1

Summary of Previous Reviews

Author, year	Topic of review	Type of sample[a]	Method of selection specified?[b]	Previous reviews cited and critiqued?	Outcomes of individual studies reported in terms of[c]	Number of studies reviewed	Conclusions about the effectiveness of treatment[d]
Fueyo, 1977	Test taking skills	Convenience	No	No	Conclusions	12	Effective
Kirkland, 1971	Test administration	Convenience	No	No	Conclusions	44	Inconclusive
Millman, Bishop, & Ebel, 1965	Test taking skills	Convenience	No	No	Conclusions	8	Effective
Roberts, 1979	Test taking skills	Convenience	No	No	Conclusions	13	Effective
Sarnacki, 1979	Test taking skills	Convenience	No	No	Conclusions	17	Effective
Sattler & Theve, 1967	Test administration	Convenience	No	No	Statistical significance	56	Effective
Vernon, 1954	Test taking skills	Comprehensive	No	No	Effect size	20	Effective

[a] If the review was based on a limited number of studies and gave no procedures for how studies were selected, it was assumed that the sample was a convenience sample.

[b] To be coded "yes," the specific procedures used to identify and select articles for the review had to be described.

[c] Effect size refers to any kind of measure which could be compared on a common metric across all studies. To be coded statistical significance, the review had to report whether the significance was in favor or against the treatment for the majority of studies reviewed. Reviews that reported the primary investigators' conclusions without mentioning statistical significance were coded conclusions.

[d] Entries in this column reflect the authors' stated opinion in the review article.

BEST COPY AVAILABLE

Two reviews contained criticism of the primary research in terms of design or confounding factors, but the consideration that reviewers gave to those problems in selecting studies, comparing effectiveness in outcomes, or drawing conclusions was undefined and appeared to be unsystematic. The results and conclusions of the review articles are discussed in the appropriate ST or TT section of this chapter.

The remainder of this chapter is divided into four sections: Reinforcement, Training Students (in test-taking skills), Training Teachers (in test administration), and Summary and Conclusions. The first two parts utilized meta-analysis procedures to integrate the existing research. Since no research was identified on the effect of training teachers to administer tests (section three), a short narrative review of research on related topics is provided. The results of the meta-analyses and related test administration research are summarized in the final section of the chapter.

Meta-Analysis of Research on Reinforcement

Research has demonstrated the positive effects of rewarding various types of academic behavior including test taking (Axelrod, 1972; Ullman & Krasner, 1965). However, no reviews of primary research were located which surveyed the studies that specifically investigated the effect of reinforcement procedures by using test score as the outcome measure.

Two recently completed dissertations (Baer, 1978; Weiss, 1980) include reviews of previous research on the effects of reinforcing intelligence test-taking behaviors. In both reviews the primary research was grouped by subject IQ level into low, average, and high. Both reviews concluded that students with initially low IQ scores show significant gains in IQ scores over controls when the correct responses are reinforced on a second test. However, studies that examined the effect of reinforcement on students with high IQs, found no significant changes in IQ from the first to the second testing. Similarly, most of the studies that examined average IQ students found nonsignificant changes in IQ levels.

The primary purpose of this section is to report the results of a meta-analysis of previous research to answer the question: Does reinforcement increase test scores? This section reviews the primary research on reinforcing test-taking behaviors and contains a description of a typical study, the results of a meta-analysis on previous studies, a summary, and the conclusions.

A Typical Study

In a typical study included in the meta-analyses, standardized tests were administered twice to two groups of students. The control group received two identical administrations, both following standardized procedures. The treatment group was given only one standardized administration, then retested using standardized procedures except that a reward was provided to students who received higher scores than they did on the first test. Test scores between the first and second test administration were compared to determine if the reinforcement resulted in significantly higher scores.

Results of the "Reinforcement" Meta-Analysis

The 41 effects sizes that were identified in 18 articles were used in the meta-analysis. A summary listing of ESs by study is included in Appendix A. The articles describe the impact of providing different types and schedules of reinforcement on the academic and IQ test scores of students aged 4 through 23. The articles from which the studies were reported, were published from 1917 through 1980. Three were doctoral dissertations.

Overall effects. Descriptive statistics were used to compare the results of reinforced and nonreinforced testing conditions. Table 2 lists the study characteristics coded for each study (including the coding categories), the number of effect sizes in each category, and the effect size. According to investigators, reinforcement was effective in increasing the test scores in 56% (23/41) of the reported effects. Only two authors concluded that reinforcement did not increase scores. Sixteen ESs were judged by the authors as being inconclusive. In other words, most investigators who have examined the effect of reinforcement on test scores have concluded that reinforcement does raise test scores.

These conclusions are empirically supported by the fact that the mean ES across all studies was .50, with a standard deviation of .58 and a standard error of .09. That is, when students are reinforced for scoring higher than predicted from the prestest, scores under the reinforced condition are one-half standard deviation higher than the mean score obtained under nonreinforced conditions. This implies that a typical student who is reinforced for scoring higher than predicted will score at the 69th percentile on an achievement test, whereas if the same

Table 2

Categories for Describing Reinforcement Studies, Number of Effects in Each Category, and Mean Effect Size

Characteristic	Coding categories	Number of effect sizes	Effect size	Standard deviations
Number of subjects	12 - 29	14	.50	.82
	30 - 100	23	.54	.43
	Over 100	4	.26	.41
Age of subjects	4 - 6	4	1.00	.56
	7 - 10	23	.37	.65
	11 - 23	14	.57	.40
IQ of subjects	43 - 85	9	1.10	.70
	86 - 100	21	.45	.43
	Over 100	11	.10	.33
Type of reinforcer	Money	5	.61	.59
	Candy	8	.54	1.06
	Praise	7	.40	.34
	Reproof	2	.10	.06
	Token	12	.55	.41
	Choice	3	.88	.20
	Prize	4	.60	.62
Type of reinforcement schedule	Immediate (after item)	31	.49	.63
	Immediate (after subtest)	6	.61	.30
	Delayed	4	.43	.62
Contingency	Contingent	32	.51	.63
	Noncontingent	9	.46	.39
Type of test	Academic	12	.53	.37
	Intelligence	29	.49	.66
Administration unit	Individual	36	.51	.61
	Group	5	.49	.54
Type of design	True experimental	23	.54	.72
	Quasi-experimental	14	.47	.35
	Pre/post	4	.37	.44
Quality of design	High	28	.46	.64
	Low	13	.58	.46
Conclusions drawn in study	Treatment worked	23	.77	.62
	Inconclusive	16	.15	.27
	Treatment did not work	2	.21	.64
	Overall	41	.50	.58

student were not reinforced, he or she would score at the 50th percentile.

For low achievers with average IQs (i.e., 85 to 115), scores at the 20th percentile under nonreinforced conditions would be at the 36th percentile under reinforced conditions. However, this translation from effect size to percentile must be interpreted in conjunction wtih findings that IQ influences the impact of reinforcement procedures (refer to IQ below for a more detailed discussion).

The Joint Dissemination Review Panel (1977) has described effects of the magnitude found in this meta-analysis (ES = .50) as educationally significant and Cohen (1977) has reported a half standard deviation as medium size. The number of effect sizes for each ES is graphed in Figure 1. ESs varied across studies, but 39% (16/41) of the studies had ESs of .50 standard deviation units or more. Nearly one-third (29%) of the studies reported larger effects (ES = .75 or higher) in favor of reinforcement.

As indicated in Figure 1, the distribution (mode = 0) of ESs is positively skewed toward high ESs. The median ES of .29 may be a better indicator of central tendency than the mean (.50) because of five extremely high ESs over 1.00. However, since three of the five ESs were from high quality studies (see Appendix A) and the median would not reflect their impact, the mean is used to represent the overall effect of reinforcement studies.

The data displayed in Table 2, show that on the average, reinforcing students for performing better on standardized educational tests results

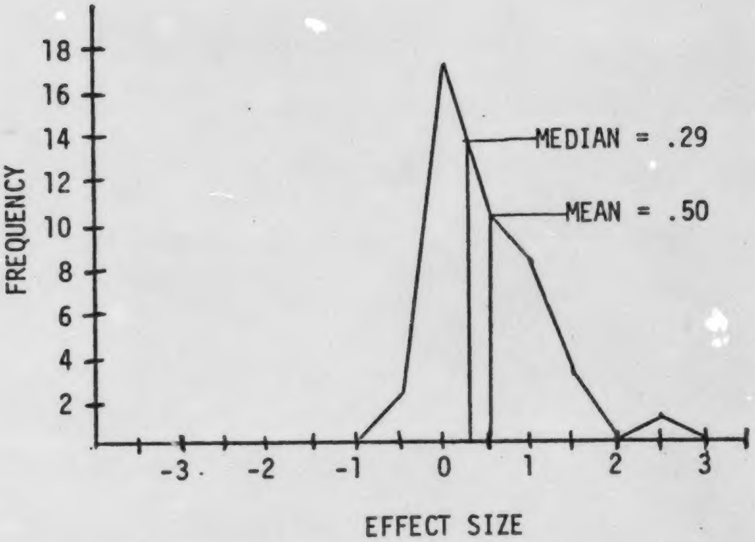


Figure 1. Distribution of 41 effect sizes for reinforcement studies considered in the meta-analysis.

in substantially higher scores ($\overline{X}_{ES}$ = .50; Md_{ES} = .28). These data imply that for some students, scores obtained under nonreinforced conditions may not be indicative of their true achievement level. However, the overall results must be interpreted in conjunction with a number of other variables considered in the meta-analysis. The most important of these variables is the IQ level of students in the sample.

IQ of students in the sample. Test scores from low IQ students (45 through 85) are more affected by reinforcement (ES = 1.1) than scores from medium (ES = .45) or high (ES = .10) IQ students. Translated into percentiles, a student with an initial IQ of 60 will receive 76.5 when reinforced on an intelligence test. A low IQ student scoring at the 20th percentile on an achievement test would shift to the 56th percentile if reinforced. An ES of .10 indicates that a high IQ student may slightly increase a score when reinforced during an intelligence test or achievement test. However, the low ES must be interpreted with caution because the student may be scoring very close to the highest possible score and may be unable to score higher regardless of motivation or circumstance.

Table 3 presents a further breakdown of how the IQ of students in conjunction with other study characteristics influences the ES of the study. Reinforced low IQ students (43 through 85) aged 4 though 6 are affected most by reinforcement procedures. Even within other categories, low IQ is associated with larger effect sizes than medium IQs which are larger than high IQs.

Table 3

Mean Effect Size by IQ for Contingency, Quality, Age, Design, Test Type, and Number of Subjects

Characteristic	IQ 43 - 85	IQ 86 - 100	IQ Over 100	Overall
Contingency				
Contingent	1.42 (5)	.48 (19)	.01 (8)	.51 (32)
Noncontingent	.70 (4)	.13 (2)	.35 (3)	.39 (9)
Quality				
High	.92 (6)	.66 (16)	.29 (6)	.46 (28)
Low	1.19 (3)	.38 (5)	-.15 (5)	.58 (13)
Age				
4 - 6	1.39 (2)	.95 (1)	.29 (1)	1.00 (4)
7 - 10	1.46 (2)	.41 (9)	.07 (9)	.37 (23)
11 - 23	.84 (5)	.45 (11)	.23 (1)	.14 (57)
Design				
True experimental	1.19 (5)	.51 (11)	.05 (6)	.54 (23)
Quasi-experimental	1.02 (2)	.35 (9)	.46 (3)	.47 (14)
Pre/post	.72 (2)	.66 (1)	.05 (2)	.37 (4)
Type type				
Achievement	.82 (2)	.48 (10)		.53 (12)
Intelligence	1.18 (7)	.42 (11)	.10 (11)	.49 (29)
Number of subjects				
12 - 29	1.24 (5)	.30 (4)	-.09 (5)	.50 (14)
30 - 100	.92 (4)	.49 (17)	.26 (2)	.54 (23)
Over 100			.26 (4)	.26 (4)
Overall	1.10 (9)	.45 (21)	.10 (11)	.50

Note. Numbers in parenthesis indicate the number of ESs.

Contingent reinforcement has a greater impact on low IQ students than noncontingent reinforcement. The data in Table 3 show that scores from low IQ students increase by 1.42 standard deviation units under contingent reinforcement conditions and are considerably more affected by reinforcement than scores from high IQ students who are reinforced contingently (ES = .01). Therefore, under contingent reinforcement conditions, a low IQ student may increase an IQ score from 60 to 81 or an achievement test score from the 20th to the 77th percentile.

In many cases the small number of ESs available for analysis requires fairly cautious interpretations of estimated impact of various conditions. However, the trend supported by these data indicates that there is an inverse relationship between student IQ and the amount of increase in test scores from unreinforced to reinforced test conditions. In addition, the data represent all of the research which could be located to address these questions and consequently represent the best estimate until further research is conducted.

Type of test and administration unit. Most of the studies measured intelligence (71%) and were individually administered (88%). There do not appear to be significant differences in outcomes between types of tests or units of administration (see Table 2), but there is evidence that group-administered IQ tests resulted in smaller effects (ES = .07, n = 3) than individually-administered IQ tests (ES = .53, n = 26; see Table 4). These data provide fairly clear evidence that reinforcement on individually-administered test results in higher scores. However, the data for group administered tests are more equivocal.

Table 4

Mean ES by Unit of Test Administration

for Type of Test

Type of test	Unit of test administration		Overall
	Individual	Group	
IQ	.53 (26)	.07 (3)	.49 (29)
Achievement	.46 (10)	1.12 (2)	.53 (12)
Overall	.51 (36)	.49 (5)	.50 (41)

Note. The numbers in parentheses indicate number of ESs.

Although, overall means indicate no differences between IQ and academic tests or individual and group-administered tests, the disparity in Table 4 between group-administered IQ and academic tests raises some important questions. The three studies which administered *group intelligence* tests were undertaken in the 1930's with elementary (ES = .08), junior high (ES = -.11), or college (ES = .23) students. The authors of those studies described the reinforcement treatment as promising prizes or providing praise and encouragement. However, rivalry appeared to be the basis for rewards and for the appeal to "try your best." In all studies, students were urged to increase their rank position by competing with those of higher standing or with the control group. The use of rivalry as a motivational technique is questionable, as demonstrated with the low mean effect size of .07. Perhaps rivalry is age-dependent; that is, it is more effective with college students than with younger students.

Two articles described the effects of reinforcing group achievement test behavior which is the focus of the present research study. In one study (Ayllon & Kelly, 1972), a classroom of 30 normal fourth-graders was given token reinforcements for correct responses to questions on a standard achievement test. Tokens were delivered after each subtest and back-up reinforcers were available after the total test. A statistically significant difference between reinforcing and nonreinforcing conditions was achieved ($t(30) = 5.90$, $p < .01$) and the effect size was .66. As with most pre/post designs, there were several factors that threatened both internal validity (history, maturation, testing) and external validity (incomplete description of treatment, Hawthorne effect, pretest sensitization). The extent to which the significant results of this experiment can be generalized is questionable due to the threats listed above, the small number of subjects, and the single classroom used. (See Campbell and Stanley, 1963; Bracht and Glass, 1968, for a thorough discussion of these rival hypotheses.)

A second study (Chapman & Feder, 1917), like many early reports, omitted much of the relevant treatment description. Essentially, extended practice on three math tests was given to two groups of 16 fifth grade students who were matched on addition test scores. Group B worked under normal conditions and Group A was given external incentives (i.e., stars and back-up reinforcers) for high scores or improvement. Data were kept for ten consecutive days and visually analyzed by daily graphing the mean test scores of both groups. The results showed the mean test score for Group B to be higher than Group A at every data point. Several methodological problems in this study threatened internal and external validity.

First, the students were all from the same classroom and were probably not isolated during the testing or the delivery of reinforcement. Therefore, the influence of prizes being given to Group B may have depressed the scores of Group A.

Second, although stars and prizes were used to motivate the students, competition was the more likely incentive for Group B. That is, each day's scores were published and students were encouraged to "beat" their last score and their classmates. Stars and prizes (given at the end of the study) were given to only the top 50% for efficiency and improvement.

The third potential extraneous variable was that students were matched on scores only from the test (addition) that obtained substantially different results between treatment and control groups. For the addition test, Group A's scores actually decreased from the first to the tenth data point while Group B's increased. Scores for Group A students increased in a similar manner to Group B in the other two tests.

Fourth, although students were matched on scores from one dependent variable, it was only a ten minute test. The fact that final scores on the other two measures did not differ between groups, creates suspicion that the matching criteria may have been biased.

Fifth, the subtests were too short (i.e., 10, 5, and 1 minutes) and not properly standardized, according to today's standards, and the number of subjects was too small to justify generalization of the results. While the data in both of these studies support the notion of reinforcement improving group academic test scores, both reports are of insufficient quality to rely on the findings.

The small number of available ESs, the poor quality of existing research on this topic, and the disparity in previous results indicate a need for additional research investigating the effect of reinforcement using group-administered tests. Research investigating group-administered academic tests is particularly important because of the frequency with which these tests are used to make educational decisions about students which might be influenced by the instructional level of the student.

Other study characteristics. IQ was found to account for most of the variance in ES across the categories of various study characteristics. To illustrate the influence of IQ, note that the data in Table 2 indicate that studies with over 100 subjects (n = 4) have smaller ESs than studies with fewer subjects. However, the subjects in those four ESs were high IQ students and, therefore, a smaller ES is to be expected.

Eighteen ESs were from studies on second grade students. Individual intelligence tests were used to measure the effect of a variety of rewards with the exception of money. Of the one achievement and four intelligence tests that were reinforced with money, all were individual exams given to fourth and fifth graders with average IQs. The most powerful reinforcer was giving the students a choice of the reward they desired (ES = .88). The least effective reinforcer was reproof (ES = .10).

Studies were coded True Experimental, Quasi-Experimental, or Pre/Post designs based on the definitions provided by Campbell and Stanley (1963). ESs were not significantly different across designs, although pre/post designs were below the mean ES (.50) at .37 standard deviation unit. When the low quality studies (32%) were removed from the analyses, the ES decreased only slightly to .46.

All pre/post designs (n = 4) were rated "low" quality. Quasi-experiments were coded "low" (n = 6) when various threats to external and internal validity were present including statistical regression, poor matching techniques, volunteers, pretest sensitization, experimenter effect, and inconsistent or poor description. Three of the 23 true experiments were coded low quality because of the use of volunteers, experimenter effects, the Hawthorne effect, and the lack of population validity.

In all the reviewed studies, "novelty" was a rival hypothesis and posed the greatest overall threat to external validity. The reinforcement procedures implemented by the investigators were always novel experiences for the subjects. That is, the treatment consisted of providing activities not typically associated with standardized testing. Consequently, differences between reinforced and nonreinforced students may be caused by experimental students attending to the newness of the reinforcement activities rather than by higher motivation to do well on tests.

Summary

The results of the meta-analysis produced substantial evidence that reinforcement techniques result in higher standardized test scores. The overall effect size of studies comparing the results of reinforcement and nonreinforcement was .50 standard deviation units. Although the median

(.29) was considerably lower, it was not used as a measure of central tendency because the effect of three high quality studies (ES = 1.98) was better represented by the mean.

A mean ES of .50 corresponds to a standardized test score increase of about 19 percentile points for typically achieving students, 16 percentile points for low achievers of average IQ, and 36 percentile points for low IQ students (ES = 1.1). Substantially smaller increases would be expected with high achieving and high IQ students.

Just over half of the effects (23/41) were from studies reporting that reinforcement was effective in increasing test scores. Ten of those used achievement tests and 13 used intelligence tests. Only five effects were from group tests (two achievement, three intelligence).

The two studies that examined the effect of reinforcement on group achievement tests had major methodological problems which prevented confident conclusions. However, the results of both studies did favor the reinforced students. Three studies that used group intelligence tests to examine the use of rivalry to "challenge" students into increasing their IQ scores had inconsistent results.

Younger students appear to be more easily influenced by reinforcement (ES = 1.00) than older students (ES =.46). All ESs from studies with second grade (n = 18) students used individual intelligence tests and ranged in ES from 2.69 to -.26.

When the poor quality studies were removed, the ES decreased only slightly from .50 to .46. The major methodological problems were the use of pre/post designs, poorly matched subjects, volunteers, and nonrandom assignment as well as violations of external validity including population validity, limited treatment description, and the Hawthorne effect. The type of design was unrelated to the magnitude of the effect size obtained.

All rewards (excluding reproof) that were investigated were effective in raising scores. Money was used as a reinforcer in five effect sizes with individual intelligence tests and was an effective agent in increasing test scores (ES = .61). No money rewards were provided with group or achievement tests.

In further study characteristic breakdowns by IQ and category, IQ was clearly the most important differentiating factor. That is, the test scores of low IQ students increased more under reinforcement than the scores of high IQ students. Furthermore, the strongest effects of reinforcement were found with young (ages 4 through 6), low IQ (45 through 85) students. These results support the conclusions reached in dissertation reviews by Baer (1978) and Weiss (1980).

Conclusions

Much research has documented that major changes in behavior rates have been produced by the application of reinforcement principles. Yet there are little data to show that these procedures can be applied to one of the most important behaviors in education: performance on group-administered standardized achievement tests.

The meta-analysis conducted with 41 studies related to the impact of reinforcement on standarized tests scores found reinforcing techniques to be effective in increasing test scores by .50 standard deviation units. Most of the previous research used individual intelligence tests and only two group achievement test studies were located. Although generalizations from only two studies must be cautiously interpreted, the mean effect size of 1.10 lends support to the notion that providing students with reinforcement will increase their group achievement test scores.

However, these studies were of questionable value because of either a pre/post design or the poor matching of a small number of subjects ($\overline{X}_N = 31$) from intact classrooms. Also the lack of treatment description makes replication impossible. No large scale, high quality true experiments have examined the effect of reinforcement on group achievement test behavior.

The fact that group testing is so prevalent in the nation's schools and that most students take at least one group achievement test per year until graduation, emphasizes the need for investigating the effect of various testing conditions on test scores. Research on the effects of providing reinforcement on student test-taking motivation during group standardized achievement testing is particularly necessary to address the following concerns.

1. The needs of students to experience highly motivating situations in all school activities including tests.
2. The elimination of motivation as an ambiguous and discriminating variable in test interpretation.

According to the meta-analysis data, an IQ score of 81 measured under reinforced conditions compares to an unreinforced IQ score of 60. For achievement tests, a reinforced percentile of 69 compares to an unreinforced percentile of 50. Since reinforcement appears to have a substantial impact on test scores, methods must be found to eliminate student motivation as a source of variance in test score comparisons.

More specifically, research is needed on the impact of reinforcement on *group* test scores. All previous research located on *primary* students has used individual testing. An examination of reinforcement techniques on the group achievement test scores of primary students is clearly an important step in furthering the understanding and interpretation of test-taking behaviors. Currently no high quality research studies demonstrate the effectiveness of reinforcing students on *group achievement* tests. Although results from *individual* testing show that reinforcement increases scores and may generalize to group testing, there are differences that should be examined.

For example, by its very nature, individual testing can encourage high student motivation due to the close proximity of the examiner and the ease of controlling undesirable effects (fatigue, illness, nervousness, and anxiety). The problems created by group testing (e.g., machine-scoreable answer forms and large group directions) are more difficult to overcome because of the large pupil/teacher ratio. Moreover, testing experiences that differ from the daily work are first encountered in the early grades.

Based on the review of previous research, there is a need for a larger scale study to investigate reinforcement procedures on test scores. Such a study should meet the following conditions:

1. *Employ a known reinforcer*. The study should not test the *strength* of the reward. Instead, the research should demonstrate the *impact on test scores* of using a known strong reinforcer.

2. Use a true experimental design. Experimental and control groups should be formed by randomly assigning whole classrooms, so that treatment conditions will be isolated from the nontreatment group. Also there will be no need to pretest for matching, thus eliminating any "pretest sensitization."

3. Specify the "treatment". Any variable that confounds with reinforcement procedures needs to be eliminated. However, the students should have experience in earning the reward before data are collected. The subjects need to believe that reinforcement is coming and know how it feels to be rewarded for some performance.

4. Use contingent, immediate reinforcement. Score improvement, not rank increase, should be reinforced to eliminate competition as an extraneous variable. The delivery of rewards based on the student's own score is more effective if reinforcement is given very soon after the test is taken.

Meta-Analysis of Research on Training Students in Test-Taking

The fact that it may be possible to raise students' test scores by training students to take tests is important since test results are used as a basis for educational decisions. For example, the limited number of slots available in some special programs (e.g., special education and Title I) requires that students score below a certain test score criteria. Additionally, test scores are important to entrance and exit requirements for college, graduate schools, and vocational institutions. Licenses for driving, specialized teaching, or practicing medicine and law are also awarded based on test scores.

Test taking is a critical survival skill in today's society. Whether this skill is learned by experience or through instruction is an issue currently facing educators. Due to the multiple choice, machine scorable answer formats of most group standardized tests, unique skills are required of students who are expected to demonstrate mastery of the information contained in the test. Among these behaviors are the elimination of obvious distractors and systematic guessing: skills which are not necessary for answering the open-ended or single response questions most frequently used in instructional settings.

This section reviews the research which has examined the effect on test scores of training students to take tests. First the test training components will be defined. Second, previous reviews on the test-taking literature will be examined. Next, two typical studies on training students will be described. The results of the meta-analysis conducted on primary research in the area will then be presented. Finally, a summary and conclusions of the meta-analysis findings will be given.

Definition of "Training"

Three types of training have been investigated by researchers concerned with the degree to which test-taking skills contribute to student test scores: practice, coaching, and training in test wiseness (TW). In this review the term training refers to any prior exposure of the students to a testing situation including any combination of the three components.

Practice. Test/retest experiences with identical, parallel, similar, or dissimilar forms have all been referred to as practice (Vernon, 1954). It is the lack of instructional feedback that distinguishes practice from coaching or training in TW. Practice is a type of "training," because it is possible for students to "teach" themselves, or learn from prior experiences.

Anastasi (1976) theorized that certain types of questions may be much easier to answer when encountered a second time. For example, some problems may require insightful solutions which can be reapplied in solving the same or similar problems on a retest. The individual who has extensive prior experience in taking tests may have an advantage in test performance over one who is taking the test for the first time (Heim & Wallace, 1949, 1950; Millman, Bishop, & Ebel, 1965; Rodger, 1936).

Coaching. Prior to the 1950's, the term "coaching" was used to describe the technique of telling students the right answers on a test and then giving them hints on how to improve their performance (Vernon, 1954). The term became synonymous with "training in TW" as it was popularized in the 1950's. In this study, *training in TW* is broader in scope and is used to incorporate all aspects of coaching as well as some form of practice on item formats.

Training in test-wiseness (TW). In recent years, the rubric "test-wiseness" (TW) has been used to describe the variables used in constructing instructional programs to teach test-taking skills. Thorndike (1951) first suggested that TW may influence the validity of a test. TW as a skill independent of content knowledge, has been defined by Millman, Bishop, and Ebel (1965) as "a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score" (p. 707).

The major skill divisions of TW have been outlined in a taxonomy by Millman et al. (1965) and include strategies for time-use, error avoidance, guessing, deductive reasoning, intent (of test constructor) consideration, and cue-use.

Rowley (1974) contends that the frequent use of multiple choice tests has precipitated the "test wise" students who receive higher scores than other students when both groups have the same knowledge. TW is not a general trait but appears to be "cue specific" (Diamond & Evans, 1972). For instance, TW students will use grammatical cues to "guess" the correct answer: a question with a plural verb form will be matched with an answer that has a plural verb rather than with an answer having a singular verb.

Certain tests are more susceptible than others. For example, TW accounted for 25% of the variance in the vocabulary test scores of ninth grade students because of the use of cues in the items (Scheib, 1979).

Novel situations, in particular, discriminate between the TW student and non-TW student (Ebel, 1976). Millman and Setijadi (1966) demonstrated that students taking a test with a familiar format do better than students who were unfamiliar with the format.

Experimental studies have shown that TW can be learned through specific training or through test-taking experience (Gibb, 1964; Moore, Schutz, & Baker, 1966; Slakter, Koehler, & Hampton, 1970). Crehan, Koehler, and Slakter (1974) found that without training, TW increases each year up to the ninth grade were statistically significant. When TW was examined over a four year period, it was found to be a stable characteristic from junior high through graduation (Crehan, Gross, Koehler, & Slakter, 1977, 1978).

Tests to support the existence of TW have been developed by Ferrell (1972) and Woodley (1975). The correlations are statistically significant between measured TW and performance on achievement tests with multiple choice items (Alker, Carlsen, & Hermann, 1969; Ferrell, 1977; Rowley, 1974) and TW and GPA (Millikin, 1976), but are not statistically significant between TW and cognitive abilities (Diamond, Ayrer, Fishman, & Green, 1976).

Ferrell (1977) argued that all students should have formal instruction in test taking to minimize the advantage test wise students have. Techniques for teaching TW have been developed by several investigators who have found that scores on TW scales consistently increase with training (Gibb, 1964). Evidence for increases in achievement test scores, however, is conflicting (Callenbach, 1973; Moore, Schultz & Baker, 1966; Oakland, 1972; Slakter, Koehler, & Hampton, 1970).

Several commercial products specifically designed to train students in TW have been marketed since 1978. Three of these training packages are Competency Tutoring Program (1979), Mini-Tests (1979), and Test Taking Skills Kit (1980). The information available from the publishers indicates that little empirical data have been collected to determine the effectiveness of the packages in teaching TW. The major problem with the research that has been conducted is that the comparison groups systematically differed in factors other than treatment implementation. The control group in all studies was formed from schools that did not "volunteer" to purchase the kits. Consequently, there may have been less concern for test-taking skills in the control schools than in the treatment schools. Schools that purchased the kits, obtained statistically significant higher test scores than those that did not.

Previous Reviews

Five review articles were located that discussed the primary research on the effect of training students in test-taking skills (Fueyo, 1977; Millman, Bishop, & Ebel, 1965; Roberts, 1979; Sarnaki, 1979; Vernon, 1954). A sixth review was located (Jensen, 1980) but not included because instead of discussing primary research, the author summarized other previous reviews. Table 1 contains a brief description of each review. The number of research studies reviewed in the five articles ranged from 8 to 20 with a mean of 14. Two articles listed one dissertation each in their references.

The review articles illustrated the common faults that were described by Glass (1977): (a) haphazard literature searches, (b) outcomes not quantified for comparisons across studies, and (c) the inappropriate use of statistical significance to integrate findings. No author (except possibly Vernon, 1954) reviewed all the literature in the field, yet the criteria for selecting articles or the method of sampling were not reported. The use of only two dissertations suggests that at least one major source of research was not searched.

Four reviewers did not quantify their findings by using a common metric to compare results across various research conditions. The statistical significance was reported only occasionally and unsystematically. However, in no case was this information used to integrate similar conclusions or to compare findings. The reviewers formed conclusions by summarizing the conclusions of the principal investigators rather than by quantifying and systematically analyzing study outcomes. Only one reviewer critically analyzed the primary studies for design and methological problems and recommended improvements.

In the earliest work, Vernon (1954).prepared the best critical review of previous research on the effect of practice and coaching on intelligence test scores. It is unfortunate that this article was completed before the majority of the primary research in the area was undertaken (1960 to 1975). Vernon also reviewed the largest number of studies (i.e., 20), thus revealing that the other articles, printed 10 to 20 years later, omitted relevant research.

To facilitate comparisons of results across studies, Vernon translated the published data from the reviewed articles into standard scores (IQ). However, the translated scores were never integrated nor analyzed for covariance with study characteristics. The major criticisms made by Vernon on the primary work that he reviewed are listed below.

1. The description of "treatment" did not distinguish between practice and coaching.

2. Most studies used pre/post designs; control groups were rarely used.

3. Researchers did not report if the treatment was conducted on identical, parallel, similar, or dissimilar forms.

The other four reviews considered research on the effect of test wiseness (TW) on standardized tests (mostly achievement tests). The authors of all five articles concurred that, in general, practice, coaching, or training in TW will increase test scores. Other conclusions drawn by the reviewers are listed below:

1. Training is not differentially influenced by age and sex.
2. Retesting with the same form results in higher scores than if a parallel form is used.
3. Certain subtest scores are more affected by training than other subtests (e.g., larger increases were found on nonverbal and spacial test items than on verbal test items).
4. Short practice exercises that immediately precede the tests are not effective in increasing scores.
5. The time between training and testing is critical (i.e., the longer the interval, the less increase in test scores).
6. Increases in test scores due to training in TW fade more quickly than increases due to practice.
7. Training in TW is more effective than practice alone.
8. TW can be acquired by students through multiple testing or taught by teachers who deliberately coach specific skills.
9. Initially, TW accumulates rapidly but a definite ceiling exists.

These conclusions must be viewed with caution since the studies included in previous reviews were neither comprehensive nor representative. Also, the findings from primary sources were summarized without systematically considering the impact of difference in study characteristics. For instance, no analyses were performed on the effect on outcomes of number of subjects, type of test administered, age of subjects, and quality of research design. Therefore, the reviewers' conclusions are based simply upon the original authors' conclusions.

In the present study, the outcome data of all studies were converted to ESs which were analyzed for covariation with study characteristics. Therefore, the summary and conclusions will not be a tabulation of the primary investigators' opinions but will result from a quantitative examination of how variables impact differently (across studies) on the outcome data.

Typical Studies

Two investigations are described to portray the most common characteristics of the studies reviewed in this section.

Practice. One group of students was administered one test on two different occasions, one week apart. The same form and level were used in both instances. The mean pretest score was compared with the mean posttest score and any increase was attributed to the effect of "practice."

Training in TW. Students were randomly assigned to experimental and control groups. Both groups were given pre and post tests. Between the tests, the experimental group was trained in skills that apply to taking exams: how to guess, fill in answer formats, eliminate distractors, and schedule time. The same test form or similar forms were administered to the students at pre and post test. The mean control group posttest score was then compared with the mean treatment group posttest score to determine if training had a positive effect (increase) on test scores.

Results of the Meta-Analysis

Sixty-two effect sizes were generated from 37 research studies which examined the effectiveness of training students in TW or providing practice on taking tests. A summary listing of ESs by study is included in Appendix A. The studies included 34 articles published from 1924 through 1979 and three dissertations completed from 1976 to 1977.

Overall effects. Descriptive statistics were used to compare the results of training students or not training students to take standardized tests. A total of 62 effects was calculated to describe the impact of training students in test-taking skills on standardized test scores. Of the 20 practice ESs, 15 concluded that treatment increases test scores and 5 concluded that it does not. The investigators of the 42 training in TW ESs concluded that the treatment worked in 31 cases, did not work in 5 cases, and was inconclusive in 6. Such rough tally seems to support the use of either practice or training in TW to obtain higher test scores but a much more thorough analysis is possible.

Across all 62 effect sizes, the mean effect size was .62 (median ES was .46) with a standard deviation of .68 and a standard error of .09. This means that, on the average, trained students scored .62 standard deviation unit (or 23 percentile points) above the untrained students on a standardized test. According to JDRP (1977), an ES of this magnitude constitutes a large gain. Cohen (1977) reports an effect size of .50 as medium and .80 as large. Therefore, .62 is indicative of quite a powerful impact.

While the majority of the ESs ranged from -.25 to .75, two thirds (40/62) were over .25. Nearly one third of the effects (18) reported a substantial ES of over .75 (see Figure 2). Although the distribution of ES is positively skewed, the median (.46), mode (.50) and mean (.62) support the central tendency for trained students to score approximately one half standard deviation unit higher than the untrained students. In this case, the median may be a better indicator of the overall ES because the nine ESs that are over 1.00 are all of low quality and they may inflate the mean (see Appendix A).

Table 5 shows the average ES for each of the study characteristics coded in the meta-analysis.

Quality and design. Low quality studies accounted for 73% of the ESs. The most common problems associated with the low quality studies were unspecified treatments, experimenter bias, and the use of pre/post designs. Treatments of practice or training in TW were inadequately defined in 65% (40/62) of the ESs. For example, it was often impossible to determine whether identical, similar, or different forms of the test were used, how long the treatment lasted, or what training components were used.

Examiner bias occurred when test administrators were aware of the experimental conditions or when the same persons conducted the practice or training in TW and also administered the test (24% of the ESs).

Studies using a pre/post design (65%) resulted in a considerably higher ESs (.77) than those using experimental designs (ES = .35). Due to inherent design problems, all pre/post studies were coded low quality and accounted for 89% of the ESs in the poor category. The internal validity of studies coded "low" was threatened because nontreatment control groups were not used with the pre/post designs and extraneous

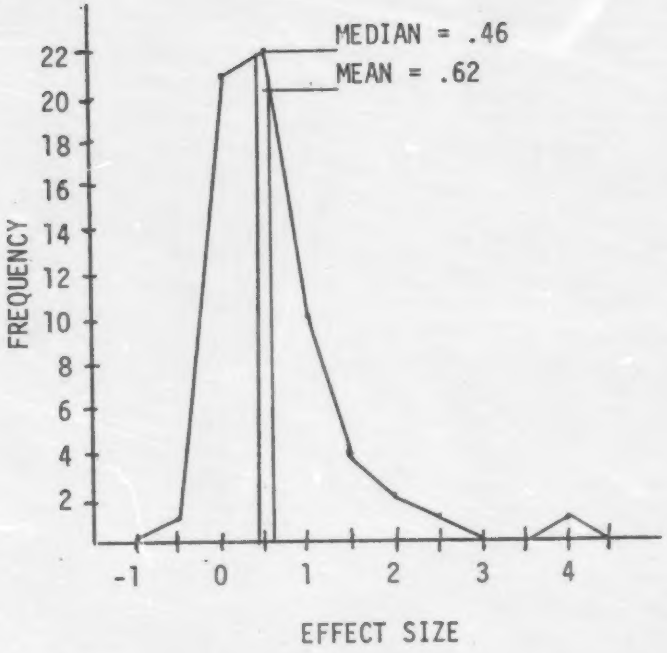


Figure 2. Distribution of 62 effect sizes from student training studies considered in the meta-analysis.

Table 5

Categories for Describing Student Training Studies, Number of Studies in Each Category, and the Mean Effect Size

Characteristic	Categories	Number of effects	Mean ES	SD
Number of subjects	9 - 49	13	1.15	1.10
	50 - 99	20	.67	.47
	100 - 199	9	.35	.32
	200 - 705	13	.37	.46
	Over 1000	7	.32	.24
Age of subjects	5 - 10	8	.69	.49
	11 - 14	22	.47	.56
	15 - 18	10	.52	.77
	19 - 24	18	.87	.83
	25 - 40	4	.44	.44
IQ of subjects	65 - 89	2	1.90	.79
	90 - 114	37	.47	.42
	115 - 120	23	.76	.89
Type of training	Practice	42	.72	.69
	Test wiseness	20	.41	.60
Type of test	Achievement	30	.40	.31
	IQ	32	.82	.85
Unit of Administration	Individual	16	1.12	.63
	Group	46	.45	.61
Design type	True experimental	17	.36	.33
	Quasi-experimental	5	.31	.39
	Pre/post	40	.77	.77
Quality of Research	High	17	.32	.31
	Low	45	.73	.67
Conclusions	Training worked	46	.78	.72
	Training did not work	10	.09	.11
	Inconclusive	6	.24	.08
	Overall	62	.62	.68

BEST COPY AVAILABLE

variables of history, maturation, and testing were not controlled (Campbell & Stanley, 1963).

When the poor quality studies (n = 45) were removed from the analysis, the mean ES became .32. True experiments accounted for 16 high quality ESs (ES = .34) and quasi-experiments for 1 ES (ES = .06). These data indicate that training is a powerful influence on test scores because even when only the best, most rigorous studies were considered, typical students will increase their scores from the 50th to the 63rd percentile after treatment.

Type of training. Studies providing practice in test taking described larger effects than studies that trained students in TW (Table 5). Some of the large impact of practice can be attributed to the 37 pre/post designs used to investigate the effect of practice (ES = .76). Thus, quality of research design, rather than the type of training, may be responsible for the difference in ES. When only the high quality studies were considered, the effect of practice (ES = .32) was similar to the effect of training in TW (ES = .33) (see Table 6).

Type of test and unit of analysis. As shown in Table 5, the 23 IQ tests administered had a higher mean ES (.82) than the 30 achievement tests (ES = .40). For most categories, IQ tests achieved a higher ES than achievement tests.

To investigate the factors that contributed to the larger effect sizes associated with IQ tests, ESs which resulted from studies with high and low quality research designs were examined separately. When the low quality ESs were removed, the effect of training on achievement tests

Table 6

Mean ES by Quality for Type of Training and Number of Subjects

Characteristic	Category	Quality of research design: High	Quality of research design: Low
Type of training	Practice	.33 (4)	.76 (38)
	TW	.32 (13)	.57 (7)
Number of subjects	9 - 49	.66 (2)	1.24 (11)
	50 - 99	.25 (6)	.84 (14)
	Over 99	.29 (9)	.42 (20)

Note. Numbers in parentheses indicate the number of ESs.

(ES = .31) was very similar to the effect on IQ tests (ES = .37, see Table 7).

Also, it is noteworthy that 84% (27/32) of the IQ test effects as opposed to 43% (13/30) of the achievement test effects, used a pre/post design. When ESs from only true experimental studies were compared, there was little difference between ESs obtained using IQ and achievement tests.

When poor quality designs were eliminated from the analysis, only three IQ test effects remained (ES = .37) and cautious interpretation is needed for so few ESs. In this group of high quality IQ test effects, practice had a lower ES (.28) than training in TW (ES = .40), whereas the overall analysis (high and low quality) on practice was found more effective than TW. Therefore, with a mean ES of .37, typical students can increase their IQ scores by 5.5 points (or 14 percentiles) with training. Only one of the three high quality IQ test effects administered the exam individually, resulting in a higher ES (.69) than the group exams (.20).

An examination of the high quality achievement test ESs in Table 7 yields only two large differences among categories. Some variance from the mean achievement test ES of .40 can be attributed to the ES of .48 of the 16 low quality designs. As shown in Table 8, 15 out of 20 aptitude test studies accounted for 90% of the low quality practice and 100% of the low quality TW effects.

The five high quality aptitude test designs used 17 to 22 year old students and all the exams were group administered. A single high quality aptitude study on the practice effect (ES = .83) was exemplary in

Table 7

Mean ES by Type of Test and Quality of Research Design for Type of Training, Unit, Age, Design, and IQ

Characteristic	Achievement tests		Intelligence tests	
	High quality	Low quality	High quality	Low quality
Type of training				
Practice	.30 (11)	.62 (10)	.28 (1)	.89 (25)
TW	.35 (3)	.28 (6)	.42 (2)	.79 (4)
Unit				
Individual		.78 (1)	.69 (1)	1.13 (13)
Group	.31 (14)	.46 (15)	.21 (2)	.67 (16)
Age				
5 - 9	.25 (3)			.95 (5)
11 - 14	.37 (3)	.39 (1)	.21 (2)	.49 (16)
15 - 18	.23 (6)	.03 (2)		1.90 (2)
19 - 24	.45 (2)	.65 (10)	.69 (1)	1.41 (6)
30 -40	1.[illegible]9 (1)	.22 (3)		
Design				
True experimental	.33 (13)	.78 (1)	.37 (3)	
Quasi-experimental	.06 (1)	.03 (2)		.73 (2)
Pre/post		.52 (13)		.89 (27)
IQ				
65 - 89				1.90 (2)
90 - 114	.42 (9)	.18 (6)	.37 (3)	.64 (19)
115 - 120	.11 (5)	.65 (10)		1.31 (8)
Type of test by quality	.31 (14)	.48 (16)	.37 (3)	.88 (29)
Type of Test	.40 (30)		.82 (32)	

Note. Numbers in parentheses indicate the number of ESs.

Table 8

Mean ES for Studies Coded "Achievement" by Quality, Test, Age, and Training

		High quality		Low qualtiy	
		Aptitude	Achievement	Aptitude	Achievement
Age 6 - 7	Practice				
	TW		.25 (3)		
Age 13 - 24	Practice	.84 (1)	.10 (2)	.64 (9)	.39 (1)
	TW	.10 (4)	.37 (3)	.33 (3)	
Age 30 - 40	Practice				
	TW		1.09 (1)	.22 (3)	

Note. Numbers in parenthese indicate the number of effect sizes.

that it was a true experimental design (a modified Solomon-four, Campbell & Stanley, 1963). The traditional experimental and control group mode leaves unanswered questions in this type of test/retest study because the treatment is the posttest. To solve this dilemma, Lucas (1972) randomly assigned Australian high school students to three groups (pretest only, posttest only, and pre/post test) and thereby controlled internal threats to validity. Although the increase in test scores due to practice was substantial, the measurement tool was a test of inference and the findings may not generalize to more typical American aptitude tests.

The training in TW used in four high quality (experimental) aptitude test ESs had little impact on the test scores. Although the difference between experimental and control groups did favor training, the impact (ES = .10) was too slight to conclude treatment effectiveness.

Nine ESs came from high quality studies using group achievement tests, two with practice and seven with training in TW. An ES of .10 when practice on achievement tests was researched, means that typical retested students would increase their percentile by 4 points, low and high achievers by 3 percentiles. With training in TW (ES = .42), the typical student increased scores 16 percentile points, low and high achievers by 13 and 10 points, respectively.

In summary, when considering the research from high quality designs, training in TW appears slightly more effective in increasing IQ or achievement test scores than practice. However, practice rather than training in TW is more effective with aptitude tests. Only one high quality effect (IQ test) was obtained from training in TW for an individual exam (ES = .69).

Group achievement tests: subjects under 8 years old. Group achievement testing often requires student responses different from those required by other school work. The new responses are particularly difficult for students encountering group tests and machine-scoreable formats for the first time. Therefore, a separate analysis was completed for the three experiments that investigated the effect of training in TW on primary grade students.

Two effects were obtained from the same study (Oakland, 1972) and represent the gain of treatment students over the control group on post test scores. Two posttests were given, one six weeks after the pretest (ES = .36) and the other six months after the pretest (ES = .15) Twelve 30-minute training sessions were taught by teachers over six weeks to a random half of the students (control students received no training). Training consisted of general test-taking skills required for readiness tests, multiple choice formats, direction vocabulary, pagination, independent work, marking answers, and left to right movement. Since the students were prereaders, no cue-related strategies were taught. The emphasis of the training was to familiarize the examinees with directions and answer formats for standardized tests. The classroom teachers administered the training and the tests, but they were not monitored during the training or the testing to ensure that the specified directions were followed. Consequently, teacher behavior during testing may have been a rival hypothesis if the test administration changed as a result of training the students.

In a second study (Callenbach, 1973), training in TW was given to a random half (n = 24) of students matched on pretest scores from two second grade classrooms. Statistical significance occurred (ES = .23) when comparing the standardized reading test scores of the treatment group with the control group. Eight 30-minute lessons were taught in four weeks by the investigator who also administered the posttest during the week following the treatment (five weeks after the pretest). Training consisted of following specific directions and using unique formats as well as time-use and guessing strategies. The effect of experimenter bias was a potential extraneous variable on this otherwise well-designed study.

The results of the two studies suggest that a month of short training sessions in TW will increase student tests scores over nontrained student scores. However, until these findings are confirmed by more research, cautious interpretation is required from only two results.

The only major methodological problem in the two studies was the failure to control for examiner effect. For instance, Oakland (1972) had the classroom teachers both train for TW and administer tests. This procedure raises questions about the influence of extraneous variables. Did the student training indirectly train the teacher more about test administration? That is, did the difference in scores come from better test taking or better test administration? Did the teacher display behaviors during the test that were reminiscent of the training sessions, thereby prompting the treatment students? Also, Callenbach (1973) trained and tested the students himself and may have biased the test administration in favor of the trained students.

In summary, several good quality studies found that group achievement test scores were higher when students were given practice and/or training in TW. In good quality studies with teenagers, higher ESs were obtained by studies using training in TW rather than practice alone. With primary students who are trained in TW, a 1/4 standard deviation unit (10 percentile points) increase was found for typical students, 8 percentile points for low achievers, and 6 for high achievers. No studies have isolated the effect of training in TW from test administration for young children.

Other characteristics. The quality of research design appears to be the most powerful differentiating variable among studies on training. Most of the variations found in Table 5 can be accounted for by the quality of the research study. For instance, there was only one substantial variation in ES as a result of the different ages of subjects. A large ES was obtained by the 19 to 24 year old group. Of those 18 ESs, 15 were from pre/post designs that investigated the effect of practice (i.e., test/retest) on test scores and 17 used college students as the subjects ($\overline{X}_{IQ}$ = 116). The fact that most of these ESs were from low quality research using subjects with higher than average IQ scores suggests that age may not be as strong a determinant of outcome as indicated by the effect sizes in Table 6. In fact, the most reasonable conclusion from these data is that age is not an important covariate in interpreting the research in TW.

At first glance, studies with fewer than 100 subjects had considerably larger effect sizes than those with more than 100. However, further breakdown (see Table 7) indicates that the ESs associated with small studies may be confounded by the quality of research. When low quality studies were removed from the ES computation, a reduction in strength occurred. The small number of available ESs from high quality studies with fewer than 50 subjects makes conclusions somewhat tentative.

The highest ESs were obtained by the studies using low IQ students (see Table 5). However, two studies used a pre/post design and were of low quality. The difference between the ESs of medium and high IQ students (after removing the low quality studies) indicates that scores from high IQ students (115 - 120) are less affected by training than scores from medium IQ students (see Table 6).

Summary

On the average, training students in test-taking skills increases test scores .62 standard deviation unit. In previously conducted research, the impact of training on test scores was demonstrated by differences in the percentile points obtained by trained and untrained students: 73 to 50 for typical students, 41 to 20 for low achievers, and 92 to 80 for high achievers.

However, a substantial contributor to the high ES of student training, was the use of low quality research design (pre/post designs). When the analyses were limited to ESs from only high quality research designs, the resulting ES was .32 (or an increase of 13 percentiles for typical students).

A further breakdown of the 62 ESs showed that training in TW was more effective than practice in increasing IQ and achievement test scores. For aptitude tests the reverse was true: practice was more effective than training in TW. The effect of training is similar on IQ and achievement tests. Small scale studies produced larger ESs than large scale studies, scores on individual IQ tests were affected more by training than group IQ tests, and higher test score increases resulted when the training materials more closely resembled the actual tests. Intensive training, close in time to the test, resulted in the highest score increases.

Two major methodological problems, other than use of pre/post designs, were identified: (a) many interventions were not adequately described and (b) examiner bias may have resulted from having the same person train and test.

Conclusions

The data provide evidence that educationally significant increases in student test scores can be obtained through practice or training in TW. The impact of training may make a considerable difference to individuals at the borderline of selection for special programs. Therefore, it is critical to understand the impact of various practice and training strategies on student test scores.

Currently, no large scale (over 100 subjects), high quality experimental studies have been conducted to determine the effect of training primary students in TW skills. Two small experimental studies which trained young students in TW reported an increase in group achievement test scores. However, external validity was threatened by the small number of subjects (less than 50 subjects; population validity) and the fact that the same person administered the training and test (examiner bias).

It is recommended that further research be conducted which adheres to the following conditions:

1) Conduct a large scale study. Students from several schools will increase the population validity.

2) Define treatment. Describe the type and amount of practice, TW components, length, and forms used, so that replication and secondary analyses can be conducted.

3) Use true experimental design. By randomly assigning students to treatment and control groups, the internal threats to validity will be reduced.

Review of Research Related to Training Teachers in Test Administration

A number of researchers have suggested that the test administrator can influence the outcome of an examination through the type of behavior he or she exhibits during testing. For instance, scores can be affected if an examiner does not follow the directions correctly. Also, negative attitudes can be subtly communicated to the students who may then perform in a less rigorous fashion (Messick & Anderson, 1970). If an examiner views the test as an imposition, an unstandardized testing situation may result since time limits may not be followed, clues or assistance may be given to students, or directions may not be given completely.

Aside from a few conceptual articles, no studies were located that directly addressed the effects on student test performance of training teachers to administer group standardized tests. Since no empirically based research was located, the studies reviewed in this section are those that are related to test administration. The studies provide background on training test administrators by demonstrating the effect of testing factors that are typically controlled by the examiner. The reviewed articles were located through the computer search that was previously described, but did not meet the criteria for a meta-analysis used in Reinforcement or Training.

Included in this review are studies that show the impact of manipulating various testing conditions surrounding student test scores or test behaviors. The testing conditions chosen for review are those that can be, and frequently are, controlled by the test examiner. The major categories of testing conditions that are controlled by the test administrator and that may vary within the realm of standardized procedures are student test anxiety level, the examiner/examinee relationship, the degree of test information given to students, the mechanics of taking tests, and environmental factors. Excluded from this review are studies which examine reinforcement or student training, which were reviewed in the previous two sections, and those that focus on analysis of the testing instrument.

Previous Reviews

Due to the paucity of research on test administration training, no previous review articles were located on the subject. However, two previous reviews did report on studies that investigated the effect of manipulating variables associated with testing. In one review, Sattler and Theve (1967) discussed the results of 56 research articles on factors affecting individually administered IQ test performance: departures from standard procedures, situational variables, experimenter variables, and subject variables. To summarize the findings, the reviewers reported the number of statistically significant results.

Although this review did not systematically analyze the articles for methodological problems, the authors stated that the most common design deficiency was the failure to use a random sample of experimenters. Four major conclusions were drawn:

1. Minor procedural changes are more likely to affect specialized groups than normal groups.

2. Children are more susceptible than college-age subjects to situational variables, especially discouragement.

3. The examiner's level of experience is not a crucial variable.

4. The subject's anxiety level is related to test performance.

These conclusions on individual testing have limited generalizability to group testing because the administration is different. For example, to test individuals, the examiners must often make subjective judgments in recording answers and scoring forms. On the other hand, group testing requires skills in maintaining control and motivating a large number of people to act as a unit. Therefore, examiner behavior will impact differently on individual testing than on group testing.

A second review discussed research concerned with the effect of testing on the students. Kirkland (1971) reviewed 44 studies that examined the degree to which situational variables impact on test scores. Studies were individually summarized in terms of increasing test scores but statistical significance was not reported. No conclusions were drawn and only studies that resulted in higher scores for the treatment group were reported in the review. A critical analysis of methodology and design was not conducted.

There is no indication that either Sattler and Theye (1967) or Kirkland (1971) reported on all the primary studies in the field or used appropriate sampling techniques to ensure representativeness. In describing the state of the research, both reviews restated the conclusions drawn from the primary investigators and made no attempt to quantify the outcome measure by converting it to a common index. Therefore, comparisons cannot be made across studies to determine the relative impact of the treatment. Additionally, neither review discussed the covariation of different study variables on the outcome.

Studies from the references of the two previous reviews were combined with those located during the computer searches to provide some background information on the training of test administrators. Since the studies represent a conglomeration of varied treatments, the research has not been integrated nor compared in the present review. Instead, this review merely describes the trends in previous research which support the use of various examiner training components. Unless otherwise stated, all of the studies reviewed found statistically significant differences in favor of the intervention.

Test Anxiety

The study of test anxiety began in 1952 with Mandler and S. Sarason's investigations into the correlations of high anxiety during examinations. The high test-anxious person attends more to self-relevant factors (e.g., the consequences of failing the examination) than to task-relevant factors (e.g., the elimination of obvious distractors on a multiple choice test, before guessing) and as a result is unable to demonstrate the extent of his or her skills or knowledge (I. Sarason, 1978; Wine, 1971).

Since a constellation of behaviors comprise test anxiety, it is difficult to document the complex condition with a single observational measure. From necessity, descriptive data on test anxiety are derived from the use of self-reports as well as simultaneous measures taken with other instruments. Self-reports consist of students responding to a single question or to a set of many questions regarding their feelings about test taking.

Using a single response item, Baird (1977) polled 4,248 college students after taking the GRE, LSAT, or MCAT, and found that 50% said they had been nervous while taking the test. Multiple response measures used to provide evidence of test anxiety are often screening devices such as the Test Anxiety Scale for Children (S. Sarason, Davidson, Lighthall, Waite, & Ruebush, 1960), Defensiveness Scale for Children (Ruebush, 1960), Inventory of Test Anxiety (Osterhouse, 1972), and Test Anxiety Scale (I. Sarason, 1978).

In a typical study designed to document the effect of test anxiety, high anxious (HA) and low anxious (LA) students are identified by using a particular screening measure. The validity of the high and low anxiety classification scheme is established by comparing the screening results with correlations of student performance on other measures.

To illustrate, in a study by Hill and Eaton (1977), the behavior of prescreened HA and LA middle-school students was observed while they worked on addition problems under time and failure pressures. HA students were found to take twice as long per problem, make three times as many errors, and cheat twice as often as LA students. However, when HA students in a related study operated under success conditions with no time limit, solutions were accurate and the pace was more rapid (Hill, 1967).

Students' scores on test anxiety scales have been correlated with scores on academic measures such as intelligence, academic, and diagnostic tests. For example, Kestenbaum and Weiner (1970) found that reading performance positively correlated with scores on achievement motivation measures, but negatively correlated with measures of test anxiety. Steininger, Johnson, and Kirts (1964) have linked high test anxiety with cheating. Data from college students questioned on attitudes about cheating revealed that students tend to feel that cheating is justified when situations are anxiety or hostility provoking. Steininger et al. concluded that tests viewed as senseless (without purpose) tended to evoke hostile, anxious feelings.

Based on the reviewed results of the studies, it appears that certain students are provoked into anxious feelings when presented with an examination. The extent of debilitation that test anxiety places on student test performance and methods for controlling anxiety are examined in the next two sections. Specifically, two questions are addressed: (a) Does test anxiety influence test scores? (b) Can test anxiety be controlled?

Does test anxiety influence test scores? The relationship between anxiety level and test performance has been investigated from two perspectives: (a) students' self-perceptions of their emotional state, and (b) observations of student behavior. Outcomes on these two measures are frequently confounded by subject selection and classification, type of treatment, and type of dependent measures. Studies that focus on the anxiety/test score relationship generally rely on self-report measures for classifying students as HA or LA and use an objective academic test as a correlate. Many researchers have found test scores to be negatively correlated with anxiety level (Alpert & Haber, 1960: Butler, 1980; I. Sarason, 1957; I. Sarason, 1963). In studies using factorial designs, research has repeatedly demonstrated that highly anxious students at all grade levels receive significantly lower test scores than low-anxiety students (McCandless & Castaneda, 1956; McCoy, 1965; Zigler, Abelson, & Seitz, 1973).

Paul and Erikson (1964) analyzed an anxiety/test score paradigm and found an interaction between anxiety and test scores. That is, a certain amount of anxiety is generally beneficial to test performance while a large amount is detrimental. When classified by anxiety level, individuals who were usually LA benefited from test conditions that aroused some anxiety, while those who were HA performed better under more relaxed conditions.

Support for anxiety as one determinant of test scores was demonstrated by Hill (1967) who examined the effects of social reinforcement given to 7-year-old students for marble sorting. The highest performance was obtained after success for reinforced LA students and after failure for reinforced HA students. Therefore, the use of reinforcement may have a differential effect on test results according to the degree of anxiety and attitude towards the test.

The negative aspects of high-test anxiety that result in low scores have been attributed to students failing to attend to relevant tasks, thinking irrelevant thoughts, and arousing emotions that interfere with performance (Alpert & Haber, 1960; Mandler & S. Sarason, 1952; Paul & Eriksen, 1964; I. Sarason, 1962).

Marlett and Watson (1968) reported that HA students spend part of testing time worrying about their performance or how others are doing, and often repeat solutions to problems. Other research has demonstrated that test-anxious students who are highly debilitative, exhibit high-pretest anxiety, poor attention, fixation on mistakes, self-criticism during testing, low academic self-perception, and no use of mental imagery during examinations (Couch, Turner, & Garber, 1979; Doffenbacher, 1978).

Nunn (1976) found a strong tendency for HA students to assign personal control to others rather than to themselves and as a result, fail to try to get high scores. Downey (1977) found that an "I can beat the test" attitude accounted for higher scores among students at similar skill levels.

The effect of previous failure and success appears to be an important explanatory variable within the HA/LA structure. In studying the academic performance of high school students, Osler (1954) observed that continual failure depressed pupil performance during examinations. Lazarus and Eriksen (1952) found that successful college students with high grade point averages (GPA) tend to have a better test performance under stress. Those with a low GPA had lower test scores under stress.

In a review of research on the relationship between test anxiety and test performance, Hill and S. Sarason (1966) concluded that highly structured testing procedures systematically underestimate the abilities and achievement of many anxious children with histories of failure in school. Even when failure has not occurred, but is a strong potential, the HA student will often falter on easy tasks (Eaton, 1979).

Can anxiety be controlled? While considerable attention has been given to determining the best strategy to use in reducing anxiety, most studies have demonstrated the effectiveness of treatment by using anxiety scales both as screening devices and as the dependent variable rather than using test scores as the outcome measure (Parker, 1980).

In typical studies, a self-report measure of test anxiety is administered to students before and after the implementation of a treatment designed to alleviate the debilitating emotional arousal brought on by an impending test. A treatment of desensitization or relaxation techniques is applied and the before and after self-report scores of treated students are compared with the scores from the control group to provide effectiveness evidence.

A study by Lent and Russell (1978) typifies the research on programs that are designed to reduce test anxiety. Prior to and following a desensitization and study skills treatment, self-report instruments and a simulated examination were administered to anxious college students. Students in the treatment group demonstrated significant improvement over students in the no treatment group on all self-report measures, but there were no differences on the academic tests. One explanation may be that the test (anagrams and digit symbols) may not have been sensitive to changes in anxiety levels. However, this theory is partially refuted by results from I. Sarason (1973) who found that LA college students perform at a higher level in solving anagrams than HA students. In considering the findings of Lent and Russell and I. Sarason, it appears as though treatment may reduce students' perception of their anxiety but does not influence the anxiety level itself nor the effect of high anxiety (i.e., low test scores).

In investigating various methods for alleviating test anxiety, researchers who have used scores as a dependent variable have found no statistically significant increase in test scores (Arnold, 1979; Friedman, 1979) even when GPA (Holroyd, 1976) or test taking-skills (Meichenbaum, 1972) have improved.

It is important to note that treatments reviewed in this section involve attacking the anxiety but not necessarily the cause of anxiety. For instance, if students are anxious because the test format is unfamiliar and relaxation techniques are provided, the test scores may not rise, but the students may be more at ease.

In a recent review of research on test anxiety, Tryon (1980) concluded from 85 studies that all treatments which reduce test anxiety are effective according to self-report instruments. However, there are conflicting results when treatments are measured by academic performance on objective tests. The most successful strategies have been those directed toward the elimination of worry through desensitization while providing study skills counseling.

Tryon (1980) located five studies using achievement tests as outcome measures, but the treatment group differed significantly in the outcome measure from the nontreatment group for only two of the studies. Four out of 12 studies found the treatment effective in reducing intelligence test anxiety.

Research design flaws may account for some of the variation in the findings of different researchers using academic tests as an outcome measure. Both Allen (1972) and Tryon (1980) reported that the quality of research design appears to be negatively correlated with treatment effect. The most common design problems found in research on test anxiety were the lack of credible, random placebo and control groups, therapist effects, the use of volunteers, and ill defined, complex, confounding treatments.

Summary. There is evidence that anxiety is associated with lower test scores. Since high anxiety students tend to have lower test scores than low anxiety students, achievement test results of HA students may be invalid indicators of academic skills or an underestimation of knowledge.

Studies which investigate ways of decreasing anxiety (and thus reducing the effect of extraneous variables that may confound test interpretations) have usually used self-report measures to demonstrate treatment effectiveness or have found statistically insignificant differences between treatment and control groups on academic test scores.

Several explanations can be offered for these findings. First, treatments currently used may not be effective in controlling anxiety; they may affect only the subjects' perception of anxiety. Second, anxiety may be reduced but students may continue to display poor test-taking skills. Third, a treatment may be so closely tied to the outcome measures that although the anxiety level is not reduced, subjects become aware of the "correct" response to make on self reports during the second administration. Fourth, the academic measures used in some studies are very short (one or two subtests) and the skill range may be to small to detect score differences due to lower anxiety levels. Finally, it may be that current treatments to lower test anxiety do not raise test scores because the underlying causes are not treated. Anxiety may result from unfamiliar test formats, strange examiners, previous failures, lack of test-taking skills, or a general misunderstanding of test directions.

Because the relationship between high anxiety and low test scores has been documented, further research is warranted to determine how to obtain measures of student achievement without the influence of anxiety. In this regard, it behooves test administrators to somehow reduce anxiety levels if students are to obtain valid, interpretable test scores. Though only indirectly associated with anxiety, some techniques have been demonstrated to be effective in raising test scores. The next sections will describe procedures that should be considered by test administrators to obtain more accurate test results. Many of these techniques may be applied within standardized conditions. The fact that these examiner behaviors are not specified in test manuals, encourages uncontrolled variation in test scores that is not attributable to differences in academic skills.

Examiner/Examinee Relationships

The importance of providing positive testing experiences is demonstrated in the literature by the low test scores which result when examiners who are strange, unfamiliar, negative, or punishing are used. Test manuals usually recommend that examiners establish rapport with students before testing, but rarely specify the procedures for establishing such a relationship. In recent years, investigators have examined the impact of examiner characteristics on test scores. For example, Masling (1960) found that test results varied systematically as a function of the examiner/examinee relationship. These differences may be related to the personal characteristics of the examiner such as sex, race, personality, or appearance (Stoneman & Gibson, 1978).

Gender of test administrator. Some researchers have shown that examiner gender influences test scores (Cieutat & Flick, 1967). One hypothesis is that elementary students are more familiar with female teachers than male and this may encourage higher test scores under female test administrators. In testing this theory, Back (1979) found in two related studies that the statistically significant high WISC scores obtained by female examiners over male examiners was reduced to nonsignificance when male examiners spent 15 minutes with the children prior to testing.

Race of test administrator. Although studies examining the effect of the examiner's race on test performance have produced conflicting evidence, the statistically significant effects found with some demonstrate that race is a potentially confounding variable (Katz, Henchy, & Allen, 1968; Katz, Roberts, & Robinson, 1965; Thomas, Hertzig, Dryman, & Fernandez, 1971). In a recent review of 16 well-designed studies on race of the test administrator, Jensen (1980) found a statistically significant interaction (race of teacher X race of student) in only six studies. Because of the inconsistency in favoring same and different race, Jensen concluded that race is not a source of test score variance.

Poise of test administrator. Even more subtle factors may influence student performance. For instance, in giving instru ions or oral problems, teachers may encourage or discourage students by the rate of speaking, tone of voice, inflection, pauses, and facial expressions (Anastasi,1976; Wickes, 1956). The examiner's behavior before and during the test administration has also been shown to affect test results. For instance, by displaying an expectation that students will perform well, examiners may create a self-fulfilling prophecy (Exner, 1966).

As early as 1949, Thorndike emphasized the importance of "presence" in a test administrator. This attribute includes assurance, poise, dominance, and a good speaking voice. To obtain and maintain control of the testing situation, Thorndike insisted that a teacher be thoroughly familiar with instructions, conscientiously follow the directions, know the principles and purposes of testing, and exercise good judgement approval group performed significantly higher than those in the disapproval group.

In a study comparing the IQ scores of students who were tested by examiners using standardized conditions, Thomas, Hertzig, Dryman, and Fernandez (1971) found that the nature of the examiner significantly influenced test results. Scores were higher when tested by a warm, friendly, encouraging examiner (who also spent more time with the students before normal testing) than with examiners who made no effort to create a positive environment.

While most studies found higher test scores associated with warm and positive test administrators, Coleman (1978) demonstrated that some types of personal interactions with teachers may be distracting to students during testing. Sixth grade students experiencing a cold, task-centered examiner did significantly better on group administered intelligence tests than students who experienced a warm, child-centered examiner.

The type of rapport existing between examiners and students prior to testing also influences test results. Emotionally or physically disturbing the examinees immediately preceding an examination significantly reduces test scores (McCarthy, 1944; Reichenberg-Hackett, 1953). Based on the premise that testing maximizes anxiety in children, Piersel, Brody, and Kratochwill (1977) found that exposing students to an affectively warm and rewarding pretest experience resulted in improved test scores and reduced apprehension levels.

Familiarity of test administrator with students. The effect of familiarity was examined in an early study by Sacks (1952). Ten-year-old students were randomly assigned to three test administrators, two familiar and one unfamiliar. One of the familiar test administrators had established a poor relationship with the students, the other a good relationship. Statistically significant differences indicated that students with familiar positive test administrators obtained higher scores than students with familiar negative examiners, who do better than students with unfamiliar examiners.

Negative prior testing experience. The effect of negative past testing experiences was investigated by Davis, Peacock, Fitzpatrick, and Mulhern (1969). Math test scores from two groups of college males were compared to determine the effect of prior failure. Those students who had failed on a previous test and had received negative feedback from the examiner performed significantly lower on the math test than students without such experiences.

Information About the Test

The degree to which examinees should be informed about the testing situation (e.g., type of test, type of test format, content, use of test results, scoring protocols, length of test, difficulty) has been debated for several decades. One perspective emphasizes the danger of instilling too much anxiety by over-emphasizing the importance of test scores in a student's future endeavors. On the other side, examinees may not try to do their best if they have not been properly informed about the test.

Advance notification. Although the effects of giving standardized tests without some sort of previous announcement has not been investigated, there is some evidence that students obtain higher scores on teacher-prepared tests if an upcoming test is announced than if it is not announced (Pease, 1930). Tyler and Chalmers (1943) found that the average scores of junior high students increased substantially by providing a specific notification that a weekly test would be given.

"Game" vs. test. The way in which tests are referred to has also been shown to affect student test behavior. For example, when third grade students in one study were told that they were to play a game, the experimental group had significantly higher IQ scores than the control group who were told that they would be given a test (Strang, Bridgeman, & Carrico, 1974). However, Orfanos (1979) found no significant difference between students taking a test or playing a game. It was concluded that the subjects, fourth and seventh grade students, were too aware of the nature of the test to be fooled into "playing a game."

How an examiner introduces a test may also differentially influence test scores depending on the students' emotional states at the time of the examination. For example, Sarason and Palola (1960) found that highly anxious college students who were told that the results of an achievement test would reflect their intelligence and predict their success in later life received lower test scores than students who were told nothing. There was no difference in the scores of low anxious students.

Knowledge of items-difficulty level. Information given to students prior to the test about the difficulty level can assist test wise students in organizing their time for a speed test. If easy and difficult items are randomly placed throughout the test, a good test taker will answer all easy questions first, skipping the unknown items for later consideration. If the questions are arranged sequentially, easy to difficult, the test-wise student will proceed through the test item by item. Kubiszyn (1979) investigated the effect of listing the difficulty level of the questions next to each item. He found that test-anxious students receive higher scores when they know the difficulty level than when they do not. It was concluded that anxious students are more relaxed in answering questions that are indicated as being "easy." In addition, in a 1978 article, Huck hypothesized that higher scores will result when students are told that an item is "difficult" because they will read more carefully than they will if an item is "easy."

Feedback on test performance. The effect of providing students with feedback on how well they are performing on tests has been disputed among researchers. In one study, giving students item by item feedback on test performance depressed the IQ scores (Piersel, Brody, & Kratochwill, 1977). On the other hand, Benson (1980) found that low ability ninth grade students who were told the correct response after each trial obtained significantly higher scores on a verbal ability test than those receiving no feedback.

Variation in the method of dispensing feedback (i.e., positive or negative) could account for the difference in results of the two studies cited above. A study by Bridgeman (1974) illustrated how certain feedback may act as a motivational variable to influence performance and create a "self-fulfilling prophecy." Three groups of seventh grade students were given success feedback, failure feedback, or no feedback after taking a scholastic aptitude test. Students given success for later consideration. If the questions are arranged sequentially, easy to difficult, the test-wise student will proceed through the test item by item. Kubiszyn (1979) investigated the effect of listing the difficulty level of the questions next to each item. He found that test-anxious students receive higher scores when they know the difficulty level than when they do not. It was concluded that anxious students are more relaxed in answering questions that are indicated as being "easy." In addition, in a 1978 article, Huck hypothesized that higher scores will result when students are told that an item is "difficult" because they will read more carefully than they will if an item is "easy."

Feedback on test performance. The effect of providing students with feedback on how well they are performing on tests has been disputed among researchers. In one study, giving students item by item feedback on test performance depressed the IQ scores (Piersel, Brody, & Kratochwill, 1977). On the other hand, Benson (1980) found that low ability ninth grade students who were told the correct response after each trial obtained significantly higher scores on a verbal ability test than those receiving no feedback.

Variation in the method of dispensing feedback (i.e., positive or negative) could account for the difference in results of the two studies cited above. A study by Bridgeman (1974) illustrated how certain feedback may act as a motivational variable to influence performance and create a "self-fulfilling prophecy." Three groups of seventh grade students were given success feedback, failure feedback, or no feedback after taking a scholastic aptitude test. Students given success feedback scored statistically significantly higher in subsequent testing than those given failure feedback.

In looking at the emotional impact of testing, Shannon (1978) examined the effect of withholding feedback from students. Findings from this investigation showed that tenth grade students who received pretest counseling or posttest score interpretation maintained the same attitude toward the subject content, whether positive or negative. However, students who received no feedback on test results had significantly more negative feelings toward the subject than the control group.

Summary. Previous research has demonstrated that the type of information given to students about their examinations influences test scores and attitudes. Although further investigations are warranted to determine the extent of the impact, test administrators must be informed that scores can vary as a result of sharing various types information. Often the test directions do not specify how to provide feedback, but previous research suggests that at the very least, students will receive higher scores if they are told of an impending test, are shown the results after scoring, and are informed about the basic test structure.

Mechanics of Test Taking

As early as 1949, Thorndike wrote that students exhibited different levels of understanding about the mechanics of test taking. Studies have shown that many students not only fail to comprehend the specified directions provided with standardized tests but also cannot make wise choices in guessing or eliminating answers (Anastasi, 1976). Part of an examiner's role is to assure that students understand the techniques for taking the particular exam being administered.

Use of separate answer sheets. Most standardized achievement tests which require specialized directions use machine scoreable answer forms. Since these forms are unlike formats of daily work encountered by students, elementary pupils are often unaware of the proper method of filling in answers for multiple choice items. In addition, Traxler (1963) found that the mean test scores from forms marked sloppily were significantly lower than scores on well marked answer sheets.

The use of answer sheets that are separated from the test booklet can be difficult for elementary students because they make mistakes as they transfer from question to answer space (e.g., marking on the wrong answer line or wrong answer space) (Bell, Hoff, & Hoyt, 1964; Cashen & Ramseyer, 1969). In one study, students in grades one to three who recorded scores on separate answer sheets received significantly lower scores than students who recorded answers in the test booklets. Even with practice, scores were lower when students used a separate answer sheet than when they answered questions in the test booklet (Ramseyer & Cashen, 1971). Similar results were found by Gaffney and Maguire (1971) that separate answer sheets from students in grades two and three were filled in improperly regardless of the directions given to the students.

Guessing and systematic elimination. Since most students complete school work on a criterion-referenced basis, they are not experienced in dealing with a situation where many answers to questions are unknown.

Therefore, it is rare that students in grades one through three, in particular, will know how to eliminate answers, guess, change answers, or check work (Erickson, 1972; Traxler, 1963). Most test manuals do not provide directions for the test administrator in teaching students to guess or check work.

Several researchers have found that guessing will raise scores regardless of the mathematical correction used in scoring (Hammerton, 1965; Sheriffs & Bommer, 1954; Slakter, 1968). Taylor (1966) and Moore, Schutz, and Baker (1966) studied the impact of using different instructions to either encourage or discourage guessing and found more omitted and unfinished items when students were told not to guess. In another study, Aiken and Williams (1978) investigated the effect of instructing students to guess and found that formulas used to "correct" test scores for guessing affect students with poor knowledge of subject matter more than those with high knowledge.

Checking work. In a related area, students frequently ask if they should change answers after reconsidering the question. Most researchers concluded that students who change answers tend to get higher scores (Berrien, 1939; Lynch & Smith, 1972; Mercer, 1979; Reile, & Briggs, 1952). Bath (1967) calculated that when a response is changed there is a three to one chance that the new response will improve rather than lower the final score. In an early report by Lowe and Crawford (1929), 21,903 true-false test items were analyzed and they found that correct changes were made almost twice as frequently as incorrect changes. Similarly, Matthews (1929) examined 22,000 multiple choice items on a college level test in which 555 changes in answers had been made. Of those changes, 52% raised the score, 21% lowered it, and 26% had no effect. On another test, 18,000 true-false items were studied and of the 570 changes, 63% raised the score, 34% lowered it, and 3% had no effect. The results of a breakdown by "superior" and "inferior" students showed that although "inferior" students made more changes, only 49% of their changes raised the score whereas 68% of the changes made by "superior" students raised the score.

This work was preceded by Lehman in 1928, who examined the results of high school students changing answers on a true-false test. He concluded that high scoring students tend to make fewer, but more correct changes, than low scoring students. Conversely, poor students often make wrong initial decisions as well as incorrect revisions. Although further research has not been undertaken to examine possible causes, Lehman suggests that low performing students may not know how to evaluate their own work.

Problem attack strategies. The procedures students use to answer questions have been shown to affect test scores. For instance, in two early studies (Holmes, 1931; Washburne, 1929), students who read the comprehension questions before reading the selection received higher scores than students who read the selection first.

In 1933, Weidemann and Newens investigated the effect of different instructions for answering true or false questions. Students were told to use a specific reasoning pattern to decide if the answer was true or false. Test scores were found to vary according to the reasoning pattern given for deciding how to answer.

Summary. Since a meta-analysis of the literature on training students with a package of test-taking skills appears in an earlier section of this chapter, only studies that examined the use of a singular TW strategy (e.g., guessing) have been discussed here. As indicated by some studies, students who guess, change answers, and use their time wisely, tend to get higher scores. The test administrator often determines if students are trained in the mechanics of test taking so that test-wise skills do not have to be a discriminating factor across students. Unless teachers are instructed to prepare students, it may not happen. Therefore, classroom test scores may be a function of test administration training, making score interpretation more difficult.

Environmental Factors

Although extensive research has not been done on the influence of various settings on group test performance, several investigations show the environment to be a potential determinant of test scores. Three studies have found that when using separate answer sheets, students sitting in chairs at tables received higher scores than students sitting in chairs with a small attached writing surface (Kelley, 1943; Traxler & Hiekert, 1942; Traxler, 1963).

The arrangement of the desks in a classroom may also indirectly impact test results as shown in a study of Fenton (1927). When college students were seated closely and thus given the opportunity to cheat, 63% of the students did cheat. In a related study, Axelrod, Hall and Tams (1972) found that when students sat in row formations, their study rates were higher than when they sat at tables. The use of row formations may also improve test performance if attentive behavior is encouraged. Environmental extremes (such as poor lighting, extreme heat, poor writing surfaces) may affect test scores. In a personal communication, Rechebei (1980) told of Micronesian students taking tests while sitting crossed-legged on floor mats. Information provided by the scoring service (Loret, 1980) indicated that some of the Micronesian scores were not valid because the pencil marks were made on tests supported by students' legs and the answers were too light to be scored by machine.

The place the test is administered may have some bearing on student scores. Seizt, Abelson, Levine, and Zigler (1975) found that IQ scores from disadvantaged preschoolers were significantly higher when they were tested at home rather than at school or in an office. In a similar study, Stoneman and Gibson (1978) found that developmentally disabled preschoolers got significantly more items correct when tested in a small testing room than when they were assessed in their own classroom.

Teachers may not be able to choose the testing setting since the classroom is often the only available place. However, the recognition that setting influences student performance may discourage the use of inappropriate places for testing (e.g., the cafetorium or the principal's office) and direct the examiner's attention to details of seating arrangements.

The atmosphere of the working situation can lower anxiety and motivation performance. Millman and Pauk (1967) suggest that students may be less anxious when they are concentrating on a task. They recommend that teachers assist the students by creating an environment conducive to concentration: quiet, separated desks, structured procedures.

Summary

Although the effect on student test scores of training test administrators has not been investigated directly, studies reviewed in this section suggest that testing conditions which are under the examiner's control do influence test scores. Students' test scores were higher when:

1. the students had low anxiety levels,
2. the examiners were familiar to the students,
3. a "positive" climate was maintained prior to and during testing,
4. the students were informed of the nature and purpose of the examination,
5. some type of feedback was given after the examination,
6. the directions and general test-taking strategies were understood by the students, and
7. an appropriate setting was used.

Since these situations are established by the test administrator, examiner behavior may be a differentiating variable in test score comparison.

Conclusions

There are no empirical studies that show the degree to which untrained or trained test administrators maintain standardized conditions or that show the differential effect of examiner training on test scores. However, each year more and more school districts (especially larger cities with evaluation units) are becoming concerned with quality control measures as they elect to supervise the testing by observing teachers give tests (Krueck, 1981).

If the conditions are thought to lower scores, school districts may provide training for teachers in test administration prior to the annual district-wide examinations. However, there are no empirical data to show that training examiners will affect test scores, will encourage the implementation of standardized procedures, will improve student test scores, or will change teacher behaviors. There is no basis for decision making on whether to provide training. Hence, decisions about teacher training are made according to budget feasibility rather than a perceived need.

Due to the need for properly administered tests, it is recommended that research on the effect of training test administrators should be conducted for three purposes:

1. to determine if training influences the implementation of standardized procedures,
2. to document the effect of training on test scores, and
3. to eliminate differences in trainers as a contaminating variable in test score comparison.

To investigate the effect of proper standardized test conditions on test scores, a true experimental study with classrooms of students randomly assigned to treatment and control groups is needed. Several outcome measures should be used to determine the influence of training test administrators: student test scores, teacher behaviors during testing, and student behaviors during testing.

Summary

Previous research has produced strong evidence that student test scores can be increased as a result of reinforcement procedures, student practice and training in test-wiseness, and manipulating various test administration techniques. Although research on the effect of systematic training of test administrators was not located, findings from studies that investigated the impact of various test administration techniques indicate that changes in variables that are under the examiner's control have a substantial effect on student scores.

Such variations in test scores have serious implications for student selection, program comparison, student diagnosis, and funding. One of the most serious consequences is that the wrong students may be identified because test scores may result in part from motivation, test-wiseness, or test administration, rather than knowledge. However, the evidence from previous research is not conclusive. Some of the previous studies have major methodological problems which raise questions about the generalizability to other students: (a) examiner bias, (b) small number of subjects, (c) no control group, (d) unspecified treatment, and (e) non-random assignment. In addition, previous research has not sufficiently investigated the effects of reinforcement and training in test-wiseness on group achievement test performance of primary aged children, or the effect of training test administrators in standardized test administration procedures.

As noted earlier, the contents of this review establish the theoretical foundation upon which the procedures and materials for this project were based. The following section describes those materials and procedures in detail. As will be noted, the rationale for what to include in much of the training materials for the experimental groups was based on the findings of this review.

CHAPTER III

PROCEDURES

As described in more detail below, participating schools from each of three districts were randomly assigned to one of three groups. Participants in Experimental Group I received all of the project's training materials (i.e., teachers were trained in standardized test administration techniques, and students viewed the How To Take Tests filmstrips, completed the workbooks, took the practice tests, and participated in the reinforcement system). Participants in Experimental Group II viewed the How To Take Tests filmstrips and took the practice tests. Participants in the control group were not exposed to any of the project-related materials. The following section contains descriptions of the various training materials which were used as a part of the experimental treatments in either groups I or II. The remainder of this chapter will describe the sample of participating schools and students, the procedures for implementing, monitoring, and assisting with the experimental treatment, and the instrumentation used to collect data about the effectiveness of the experimental treatments.

Description of Materials

Based on the review of literature reported in Chapter II and the results of the previous State Refinements contract, the following four areas were identified that might adversely affect the validity of students' scores on standardized achievement tests.

1) Differential levels of test-taking skills on the part of students.
2) Students' lack of familiarity with and consequent confusion from the question format used in the district's standardized test.
3) Lack of motivation on the part of students to do their best on the standardized test.
4) Inappropriate administration of the standardized test.

The materials described below were developed by the project to eliminate or substantially reduce the influence of these variables on students' standardized achievement test scores.

Filmstrips and Workbooks: Teaching Students How to Take Tests

As noted in the review of literature, previous research has demonstrated that training students in test-taking skills raises the students' scores on standardized tests. The fact that students' scores on a test of reading comprehension can be raised by training them in test-taking skills suggests that some factor besides reading ability is being measured by the test. Since students already possess test-taking skills to different degrees, a training program which will allow all students to master test-taking skills will increase the validity of the test for measuring reading comprehension. This increase in validity results from the fact that once all students have mastered test-taking skills, the skills are no longer differentially affecting or confounding scores on the test. The student training materials used in this project consisted of nine instructional filmstrips, nine tape-recorded narrations, and accompanying student booklets. The development and content of these instructional materials are described below.

Development of training objectives. In developing the training materials for teaching test-taking skills, an analysis of the content, directions, and format of frequently used standardized achievement tests served as the primary resource. To decide which standardized tests should be examined, information was considered from the following sources: (a) which tests are used by Title I projects in Utah, (b) which tests are used by Title I projects nationally, (c) which tests have been formally adopted by districts and states, and (d) which tests were being used by the districts willing to participate in the project.

The number of Title I projects in the state of Utah utilizing a particular standardized test is shown on Table 9. Tests used by districts participating in the project are noted with an "*".

Table 9

Use of Tests in Utah Title I Projects

Number of Title I projects	Test
9	California Achievement Test
9	Gates-McGinite
8	Stanford Achievement Test*
5	Iowa Test of Basic Skills*
4	SRA
3	Woodcock Reading Test
2	Comprehensive Tests of Basic Skills*
2	Metropolitan Achievement Test*

* indicates a test used by a district participating in the project.

The frequency of use of a particular test by Utah Title I projects was somewhat different than the frequency of use by all Title I projects in the country. According to staff at the Northwest Regional Educational Laboratory (NWREL), national project utilization of tests occurs in the following order: CAT, SRA, MAT, Gates-McGinite, SAT, ITBS (see Appendix B for letter).

Staff at NWREL also reported frequencies indicating test adoptions for both districts and states by region as reported by McGraw-Hill. (Note: This information should be interpreted cautiously since it was part of McGraw-Hill's promotional material.) Table 10 displays the district and state adoption totals by region (see Appendix B for a complete listing).

Table 10

Number of Test Adoptions for Districts and States by Region

	CAT	CTBS	ITBS	MAT	SRA	SAT
Midcontinent region						
Districts	9	1	9	2	1	
States		2	1			
Western						
Districts	11	12	1	1		1
States	2	2				1
Southern						
Districts	10	10	3		8	3
States	5	1				
Eastern						
Districts	10	2		1	5	
States	2	1		1	1	
Total						
Districts	40	25	13	4	14	4
States	9	6	1	1	1	1

Using the preceding information, decisions were made about which tests to analyze in developing the student training materials for taking tests. Table 11 summarizes the rationale for the six tests included for analysis. Each of the tests listed in Table 11 was analyzed to identify (a) difficult vocabulary, (b) difficult phrases, (c) series of directions, (d) new symbols, and (e) examples of different response formats. An example of the data collection form used to analyze tests (this particular form was for the reading comprehension subtest of the MAT) is included in Appendix B to illustrate the type of information obtained. Similar analyses were completed for each test. In addition, as shown in Table 12, each test was examined to determine which subtests were included in the total reading score, the number of items in each subtest, the minutes allowed for each subtest, the content of

Table 11

Summary of Test Use for Project Tests

Test	Description of Utilization
CAT	- Most commonly used by Title I projects in Utah and nationally. - Commonly used in all regions. - Most often adopted by districts and states.
CTBS	- Not commonly used by Title I projects in Utah or nationally. - Adopted by many districts and states, especially in the West and South. - Used by Cache School District.
ITBS	- Used by 5 Title I projects in Utah but seldom used by Title I projects on a national level. - Adopted by districts and states primarily in the Midcontinent region. - Used by Nebo School District.
MAT	- Used by only 2 Title I projects in Utah, but third most often used nationally. - Adopted by few districts and states. - Used by Logan School District.
SAT	- Commonly used by Title I projects in Utah, but not nationally. - Seldom adopted by districts or states. - Used by Granite School District.
SRA	- Commonly used by Title I projects nationally and by 4 projects in Utah. - Adopted primarily by Southern and Eastern districts. - Used by Alpine School District.

each subtest, and the format for administering each subtest. The contents of the subtests making up the total reading score for each test were similar. However, several subtests were unique only to one test (SAT, Reading: Part A; ITBS, Sentences, Word Analysis; MAT, Word Knowledge; SRA, Listening Comprehension).

Based on the analyses described above, test-taking skills to be taught during the student training were identified and phrased as objectives. The

Table 12

Subtests

Test	Level	T or S[1]	Category[2]	Tests in "Total Reading" Score	Number of		Related Subtests
					Items	Minutes	
CAT	12	T			10		
		S	W	Phonics Analysis	15	25	
		S	W	Structural Analysis	11	14	
		S	V	Reading Vocabulary	15	13	
		S	PC	Reading Comprehension	20	20	
CTBS	C	T	V	Reading Vocabulary	33	15	
		S	SC	Reading Comp.--Sentences	23	20	
		S	PC	Reading Comp.--Passages	18	21	
ITBS	8	S	SC	Pictures	23	12	
		S	SC	Sentences	16	7	
		S	PC	Stories	28	15	
		S	B,V		30	14	Vocabulary
		S	W		57	20	Word Analysis
MAT (71)	P2	S	V	Word Knowledge	40	18	
		T	W	Word Analysis	35	15	
		S	SC	Reading--Sentences	13	7	
		S	PC	Reading--Stories	31	23	
SAT	P2	S	B	Reading Part A	45	20	
		S	PC	Reading Part B	48	25	
		T			30	10	
		S	W	Word Study Skills	35	15	
		T	V		37	20	Vocabulary
SRA	C	T	W	Letters/Sounds	20	15	
		T	-	Listening Comprehension	20	25	
		T	V	Vocabulary	25	15	
		S	PC	Comprehension	24	30	

[1] Teacher directed (T) or Student directed (S)

[2] V - Vocabulary
W - Word Analysis
B - Both Vocabulary and Word Analysis
C - Comprehension

original list of objectives was too long. Given the limited amount of instructional time (approximately 270 minutes) available for the student training, the original list of objectives was reduced to include only those skills which were needed most frequently across the six tests. The tests were again analyzed, the most frequently occurring skill areas were identified, and objectives for nine 30-45 minute instructional lessons were finalized (see Table 13). Skill areas making up the nine lessons included both general

BEST COPY AVAILABLE

Table 13

Objectives for Student Training Filmstrips

FILMSTRIP 1--INTRODUCTION TO FILMSTRIP SERIES

1. Understand that it is important to listen carefully and try your best on tests.
2. Start working at "go" sound.
3. Stop working at "stop" sound.
4. Put finger on page or item number when directed to do so.
5. Follow one-step directions in the booklet.
6. Stop working when the stop signal is given before a task is finished.
7. Work fast when told to do so.

FILMSTRIP 2--MECHANICS OF TEST FORMAT

1. Understand that test scores are used to determine what students need to learn.
2. Mark only one answer for each question.
3. Use answer space, circle, and oval interchangeably.
4. Mark answer space correctly.
5. Erase completely.
6. Work a "sample" with the class.
7. Follow four-step directions.
8. Work items in sequence whether items are arranged in rows or columns.

FILMSTRIP 3--RULES FOR TAKING TESTS

1. Raise their hands if they need a new pencil or if they need help from the teacher.
2. Understand that the teacher may help with directions but may not help figure out answers.
3. Point to every word as they read the test item.
4. Stop working when they see a stop sign.
5. Go on to the next page when they see a "go on" sign or if nothing is printed.
7. Go back and check their work.

FILMSTRIP 4--VOCABULARY I

1. Tell what a vocabulary test is.
2. Find a word that means the same as an underlined word.
3. Tell if the right answer names the whole picture, names part of the picture, or tells about the picture.
4. Tell why a "tricky" answer is wrong.
5. Use clue words to find word meanings.
6. Substitute printed clue words with answer choices.

FILMSTRIP 5--VOCABULARY II

1. Find the word that is opposite of an underlined word.
2. Find a word that means the same as a definition given orally.
3. Tell why tricky answers are wrong.

Table 13 (continued)

FILMSTRIP 6--WORD ANALYSIS

1. Find the letters that stand for the beginning or ending sound in a word.
2. Find the letters that stand for the middle vowel sound in a word.
3. Find the word with the same sound as a spoken word.
4. Find the word with the same sound as the underlined letters in a written word.

FILMSTRIP 7--TEST-TAKING STRATEGIES

1. Select one answer for each item for three-item pictures.
2. Check three-item pictures by seeing if all answers relate to each other.
3. Find the best word to describe a picture.
4. Discriminate between tricky wrong answers that are look-alikes and relatives.
5. Use the information in the picture to find the right answer and not be swayed by personal experiences.
6. Eliminate obvious wrong answers and then guess.

FILMSTRIP 8--SENTENCE COMPREHENSION

1. Do sentence comprehension test items in three formats.

    a. Find sentence that tells about a picture.
    b. Find word that completes sentence so it tells about a picture.
    c. Find word to complete a sentence so that it makes sense.

2. Tell why an answer choice does not make a true sentence.
3. Try each answer choice in a sentence before marking the correct word.

FILMSTRIP 9--PARAGRAPH COMPREHENSION

1. Find the answers to literal comprehension questions.
2. Find the answers to inferential comprehension questions.
3. Find the answer that tells the main idea of a story.
4. Find the best name for a story.
5. Tell why distracting answer choices are wrong.

test-taking skills, as listed under Filmstrip 2--Mechanics of Test Format and Filmstrip 3--Direction Following; and test-taking skills specific to subtests of the six standardized tests analyzed such as those reflected in the objectives for Filmstrips 3-9.

Skills which are general to all standardized tests (such as marking an answer space, erasing, working a sample, stopping and checking work) are the first skills taught (Filmstrips 2 and 3 in Table 13). Skills specific to subtests were taught next in a sequence which moved from simple to complex (e.g., simply finding the word that best tells about a picture, to finding the main idea of a paragraph). Prior to the instructional lessons on general and specific test-taking skills, students were taught how to respond to the medium of instruction used in the training package (see Table 13 for skills listed under Filmstrip 1--Introduction to Filmstrip Series).

Rationale for filmstrips as the medium of instruction. Several alternatives were considered for delivering the content of the student training (e.g., classroom teacher lecture, staff presentation, student workbooks). A major concern with most approaches was that consistency across classrooms would be difficult to maintain. Instruction provided by classroom teachers or project staff would probably vary in quality from classroom to classroom and threaten the internal validity of the study. Another concern was the amount of time required for teacher preparation. If teachers had been asked to prepare for nine 30-minute presentations, it would probably have required at least 270 minutes of preparation time per teacher (30 minutes for each lesson). By using filmstrips as the medium for implementing instruction, we could be more confident that the entire treatment was being implemented and that the quality of instruction was consistent in each of the 40 classrooms. In addition, teacher preparation time was reduced to 90 minutes per teacher (10 minutes for each filmstrip).

The filmstrips were developed to be shown on a classroom chalk board. The characters and pictures are line drawings which appear in a chalk color on the board. Classroom students were surprised and intrigued by the realism of the filmstrips--almost as if large characters drawn on the chalkboard had come to life. The fact that the filmstrips were so different from anything students had seen before helped to keep their attention, and the simplicity of the line drawings and chalk color helped to maintain the students' attention on the instructional content.

Instructional philosophy. The material in the nine filmstrips is taught using a "direct instructional" format. That is, specific skills are modeled, then the students are guided through practice and are tested on their competence. The *direct instructional* sequence is used (a) to clearly establish the intent of the instruction, (b) to reduce incorrect responses, and (c) to provide students frequent opportunities to practice and to provide the teacher with frequent opportunities to determine how well the students are progressing.

The five types of instructional objectives in the direct instructional method are listed and defined below:

Objectives	Key Words
1. *Teaching Objective*	
The students are told the specific task to be learned.	"You will learn . . . ."
2. *Modeling Objective*	
The correct way to complete a task is demonstrated. Non-examples of the task may also be shown.	"This is the right way. . . ." "This is not . . . ."
3. *Leading Objective*	
The students respond with the filmstrip characters or the teacher.	"Say it with me. . . ." "Do it with your teacher. . . ."

4. Testing Objective

The students respond alone.	"When I say go, you . . . ."

5. Correcting Objective

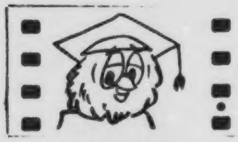
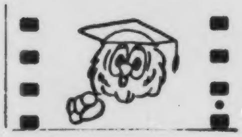
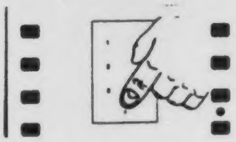
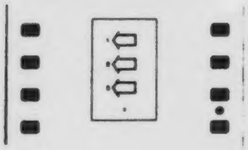
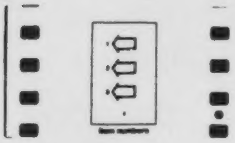
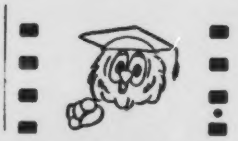
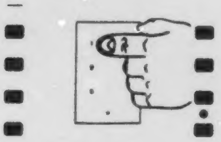
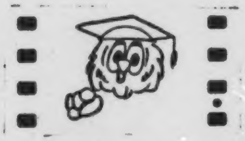
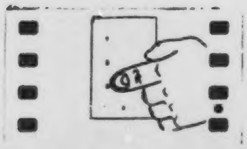
The filmstrip or the teacher show the correct response.	"Your answer should look like like this. . . ."

An example of how the five-step instructional sequence is used in the first filmstrip is shown on the next page.

Use of story line and characters. After selecting filmstrips as the instructional medium for training students in test-taking, the next step was the development of a story line and characters. Several exciting scripts with amusing and involved plots were written and piloted with individual children. During the pilot testing of these scripts, staff noticed that the complexity and interest of the story line was interfering with students' ability to attend to the instruction. Consequently, we decided that the story line must be kept simple--enough to be of interest but not so interesting that it would interfere with instruction.

Familiar animals with typical distinguishing characteristics and predictable personalities were chosen as the main characters (e.g., the wise owl, the smart and crafty fox, the slow and lovable gorilla, and the shy raccoon). The characters encourage students by stressing the importance of learning to be good test-takers. In Filmstrip #1, Professor Owl tells his animal class, "Did you know that there are magic tricks to taking tests that everybody can learn? Yes, indeed." They also offer helpful "hints" or learning strategies. For example, in Filmstrip 6 students are told by Professor Owl, "Here is a hint. You first say the word and the sound. Let's pretend the word is cat. You say the word and the sound like this. Cat--k." Throughout the filmstrip series, the characters offer timely prompts (e.g., erase completely, be sure to check your work, don't try to find the same letters).

EXAMPLE OF DIRECT INSTRUCTION SEQUENCE

Instructional Sequence	Video	Narration
TEACH:		Now, it's time to learn a new word. We will learn it on the next page, but you must listen very carefully.
		Point to page number three. Your finger should be pointing to the number three . . .
		. . . at the bottom of page three, like this. Listen, here comes the new word.
MODEL:		These three numbers are called . . .
LEAD:		. . . "item numbers." What are the numbers called? Item numbers.
TEST:		Good. Now, point to item number one. You should be pointing to item number one . . .
CORRECT:		. . . like this.
TEST:		Now point to item number three. You should be pointing to . . .
CORRECT:		. . . item number three, like this. Good.

The characters also reinforce students for trying and for learning. In Filmstrip #6, Owl says, "You did just fine, Racky. And so did everybody else. I'm proud of you." And in Filmstrip #2, Owl remarks, "You students were really, really good," and the rest of the characters respond, "Yeh, they are fantastic."

Throughout the nine filmstrips, the characters demonstrate some of the anxieties which students may be feeling about test-taking. For example, in Filmstrip #1, Owl announces that the animals are going to study a very interesting and important subject--how to take tests. The characters respond as follows:

Everyone: [Gasp] Tests!
Owl: Of course! Don't you like tests?
Gorilla: Not me!
Bunny: Me neither.
Mice: Neither do we!
Foxey: Well, I do!
Everyone: Booooooo!!!!
Owl: Now, now, students. Just a moment, please! Being able to take tests is very important to learn!
Racky: But taking tests always scares me. I mean, I just get t-e-r-r-i-f-i-e-d!!!
Gaffy: S-s-so do I.
Gorilla: Even I get scared.
Foxey: Well, not me!

Owl put all the characters (and hopefully any students also anxious about test-taking) at ease by telling them that taking tests is easy for Foxey because he knows the secret. Owl explains that there are magic tricks to taking tests that everybody can learn.

Characters also point out misconceptions and model correct and incorrect test-taking strategies. In Filmstrip #5, Simon the Snake tries to trick students into selecting the wrong option for the following item:

"Huge" means

O laugh
O hug
O large
O small

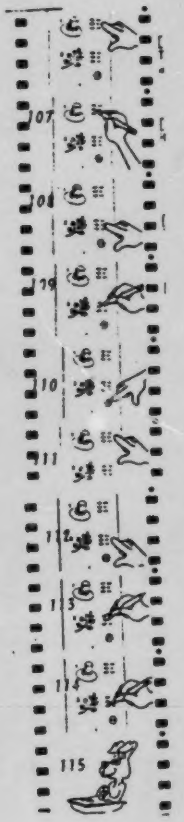
Simon makes these comments:

> - 'Laugh' isss a good choicccce. 'Laugh' looksss a lot like 'large,' 'hug' would be a sssensssational anssswer because 'hug' looksss lotsss like 'hug,' and 'small' could be a good anssswer because 'sssmall' is the opposite of 'huge.'

Each of these false lines of thinking is corrected by Owl and classroom students.

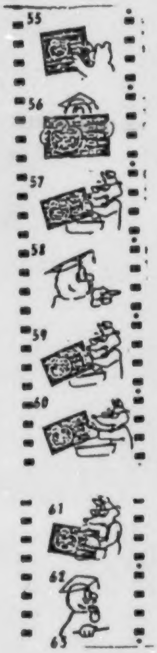
Characters also take turns modeling correct implementation of test-taking skills. In Filmstrip #3, Jack models "checking your work" by completing several test items and verbalizing his thoughts:

> OK, let's see. First, I look at the picture. Then I point to each choice as I read it. Head, hat, wear.
>
> Hat is the correct choice, so I mark hat. Ho huh!
>
> Then, I read the second item. Tree, fall, leaf.
>
> Then I mark fall.
>
> Now, I've come to a stop sign. The teacher hasn't said "stop" yet, but I am finished with my items, ho huh. So, I will go back and check my work.
>
> Hat. Yup, I still think this is the best answer so I'll leave it the way it is.
>
> Fall. Ooops, that's not the best answer.
>
> I'd better change it now.
>
> There! Ho huh! That's better. The teacher says "stop," . . .
>
> . . .so I don't have time to do any more. I just put down my pencil and wait to hear what I am supposed to do next.



In Filmstrip #3 Foxey, who is usually right and always stuffy, provides an example of an incorrect test-taking strategy. A test-taking rule (point to every word in a test item as you read it and think about it before you pick your answer choice) has just been given by Owl. The following sequence then takes place:

> Here is a test item. Foxey, show us how to follow rule number three for this test item. Point to each word as you read it before you tell us the best answer.
>
> Oh, don't be so stuffy, Professor. I don't have to point and read every word. I can tell with just a quick glance that the answer is "food."
>
> Now, Foxey, don't answer too quickly. The rule is, "Point to every word in the test item as you read it and think about it before you pick your answer."
>
> Ho hum. What a bore! Ok, food, . . .
>
> dog, . . .
>
> Oh, my, apple! Why, that *is* a better answer.
>
> See there, Foxey? That is why it is important to point to every word as you read it.



Professor Owl is the most prominent character in each of the filmstrips . . . always wise, honest, straightforward, and the primary teacher. Each of the nine filmstrips has a simple central theme. The characters interact with each other just enough to add interest and develop the outlined themes. One or two characters are prominent in each filmstrip, with new characters such as Detective Nancy True and Erp occasionally emerging.

Teacher/filmstrip interaction. The filmstrips are constructed so that the teacher must interact with the filmstrip characters and with the classroom students. Several different teacher response modes are included. For example, Professor Owl asks the teachers to *answer short questions* (e.g., "Teacher, how did your students do?"), *explain or review concepts* (e.g., "Teacher, could your students explain this so Racky will understand that trying hard is important?"), demonstrate skills (e.g., "Teacher, would you demonstrate how this page should look?"), and check the students' work and report back (e.g., "Teacher, would you check with your students to see if they answered the items correctly?).

The rationale for involving the classroom teacher was to improve the quality and flexibility of instruction. The teacher performs many tasks throughout the nine filmstrips which otherwise would be difficult, if not impossible. For example, the classroom teacher:

1. Reviews important objectives of test-taking.
2. Demonstrates continuous hand movements that are difficult to convey in still picture frames (i.e., the correct way to quickly fill in an answer space).
3. Monitors student responses, reinforcing correct responding and stopping the filmstrip to correct errors.
4. Provides a prompt (hand signal) for students, cuing them when to respond.
5. Demonstrates complicated procedures that require several steps.
6. Leads and corrects practice exercises in student booklets that reinforce skills taught in filmstrip.

The interaction between the filmstrip characters and the classroom teacher also provides diversity and maintains student interest. As classroom teachers became actively involved in the student training, the students seemed to sense the importance of the material. The teachers provided excellent models, and the students strove to please the filmstrip characters and the classroom teacher. Each time a teacher response is required, Professor Owl addresses the teacher directly (e.g., "Teacher, would you and your class like to join us to learn these magic tricks about test-taking?"). At the end of the question or request, there is a signal to the teacher and a brief pause (or blank spot) on the tape (approximately 2-3 seconds) which allows enough time for the teacher to give a short response. Rather than trying to estimate how long the teacher's response would take each time and pausing the tape accordingly, all pauses are a standard length. If the teacher wants to do more than can be done in the 2-3 second pause, he/she can turn off the tape and take as much time as needed. In this way, the teacher retains complete control of the instruction and can adjust the pace and emphasis to suit the needs of individual students. Teachers were also encouraged to circulate about the room during the filmstrip to check students' work and reinforce good behavior.

Cue cards are also provided with each of the nine filmstrips. These cards illustrate main points from the filmstrips. The purpose of these cards is to provide a technique by which the classroom teacher can easily review these main ideas. Prior to showing a filmstrip, the teacher's guide directs the classroom teacher to review main points, using the cue cards provided.

Student response mode. Throughout the filmstrip, students are asked to respond as a group either verbally, physically, or in writing. Filmstrip #1 (Introduction to Filmstrip Series) teaches the skills students need to appropriately interact with the filmstrip. Group response is used to keep all students actively involved in the learning process as well as to provide feedback to the teacher on the level of student skill acquisition. By involving the students in group response activities, the teacher can quickly survey the class to determine who is following the lesson and who needs special attention.

When a student response is requested, a question is asked by Professor Owl as he looks at the classroom. Oral responses are followed by a correction statement (i.e., "Yes, you can erase, but do not erase too often"). Classroom teachers were encouraged to elicit a response from every student because a few non-responders (who may not need to answer to learn) model inappropriate behavior for those who do need to respond. It was suggested to the classroom teachers to provide a quick drop of the hand or snap of the fingers as a signal to the students to respond. If all members of the class are responding, an active, exciting learning environment is generated, attention is kept to the task at hand, and off-task behavior is not a problem. As a rule of thumb, students were given at least two examples as part of an instructional sequence which provide students verbal practice before requiring any written responding.

The most common physical response is pointing to a page or item number. Here Professor Owl tells the students exactly where to point, and students were prompted to follow these directions explicitly for two reasons. First, pointing to things is an important test-taking skill for young students; it helps them keep their place and forces them to read every word. Second, if students are pointing, a teacher can quickly scan the desks of every child at any time and see if everyone is on the right page or item.

When a written response is required from students, Professor Owl orally signals a "go" and "stop." All written tasks are performed in a student booklet within a time limit to give the students practice in concentrating on their task and working as quickly as possible. Also, the time limit keeps the students moving as a unit which is a requirement for group testing.

Individual student work booklets accompany each of the nine filmstrips. These booklets provide short exercises so that students can practice the skills presented in the filmstrip. The booklet exercises are short, with either the filmstrip characters or the classroom teacher leading the instruction. The booklets allow the students an opportunity to practice and correct newly learned concepts before proceeding to learn additional concepts.

The format of the student booklet items is representative of the various formats used in the six tests analyzed. For example, the format of items one and two in booklet #6 is as follows:

1. g j f
0 0 0

2. ch k f
0 0 0

The answer spaces are arranged horizontally rather than vertically and the letters are above the answer spaces rather than below. This item format is representative of formats commonly used in the six tests analyzed. The content of the items is a natural continuation of the instructional examples used for modeling and leading within the filmstrip. Some of the booklet items are completed with the Owl or other characters, and some of the items are completed independently by the students.

To facilitate the development of class group response, the teacher is also encouraged to employ group response techniques when doing other activities related to the filmstrip (e.g., reviewing previous lessons, reteaching confusing concepts, asking questions, warming up the class before showing filmstrip). A hand drop or finger click is a useful cue to students that an oral response is requested.

Field tests and pilots. Instructional sequences for each filmstrip were field tested with individuals and small groups of students before the story line was added. One staff member acted as the teacher and used cue cards as a visual stimulus to walk one to three students through the entire instructional sequence (including the student workbook). This filmstrip simulation was observed by other staff members who noted instructional or procedural errors, such as the omission of proper verbal signals (e.g., "When I say go, mark your answer space") or a simple rewording which added clarity. If the students were unable to respond appropriately, one of the following changes was usually needed: (a) more modeling or leading, (b) a helpful "hint", (c) a prompt, (d) a visual stimulus (underlining of key words or character pointing to key words), (e) addition of a simpler lead-in task, or (f) a re-evaluation of objectives. Following this pilot, corrections were made and another pilot was conducted before the story line was added. For some filmstrips, the cycle of pilot-revise-pilot was repeated several times before adding the story line.

After the story line was added, the finalized script was put into story board form and photographed. Slides were then produced and sequenced in trays to pilot test before a filmstrip was produced. Pilot tests of the slides were conducted in one or more of the four pilot classrooms of Logan School District. One staff member served as the classroom teacher, one operated the slide projector, and several observed, taking notes. Following the pilot test with slides, staff members discussed their notes and decided on specific changes to be made. Because of the extensive field testing prior to pilot testing with the slides, most of the corrections which needed to be made with the slides at this point were minor (e.g., enlarge the print, eliminate red highlighting, add more character prompting). Necessary corrections were then made (i.e., new slides), and a second pilot was carried out if changes were substantial. The filmstrip was then produced, the accompanying tape was finalized, and duplicate copies were made.

Sequence of making a typical filmstrip and tape. There were many steps in making a typical filmstrip, with the tasks moving from one staff member to another and, in some instances, small groups working together.

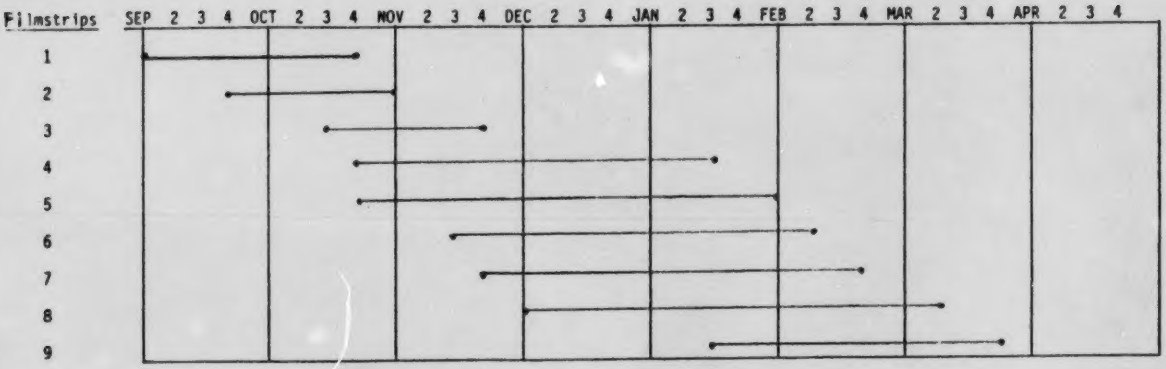
To summarize the activities involved in producing the student training materials, the steps in making a typical filmstrip and tape are outlined in sequential order below.

1. Write the instructional sequence based on the objectives for that filmstrip.
2. Develop student booklet along with the instructional sequence.
3. Field test the instructional sequence with one to three students.
4. Revise the instructional sequence.
5. Repeat steps 3 and 4 as necessary.
6. Write the story line.
7. Do artwork and photograph slides.
8. Produce a tape.
9. Pilot test the slides and tapes.
10. Revise script and retape.
11. Correct slides as necessary.
12. Repeat steps 9 to 11 as necessary.
13. Produce the filmstrip.
14. Redo the tape incorporating the corrections.
15. Make duplicate copies of the filmstrip and the tape.

The time required to complete a filmstrip varied greatly. As might be expected, the first filmstrips and tapes required more time to make because of the unfamiliarity of tasks required. With later filmstrips, the time required to produce a filmstrip and tape decreased. Table 14 shows the approximate timelines for making the filmstrips and tapes.

Table 14

Time Line for Making Filmstrips and Tapes



Practice Tests

Past research has shown that the following conditions are associated with increased student scores on standardized tests: (a) administering practice tests prior to the actual test, (b) using practice forms that closely resemble actual test forms, (c) giving feedback to students on their test performance, (d) training students to work independently for up to 30 minutes, (e) giving students timed tests in reading and math prior to the actual test, and (f) familiarizing students with the directions. As a result of these research findings, the use of student practice tests was incorporated as an integral part of the present project.

Students in both the Experimental I and Experimental II groups were provided with practice in taking standardized tests throughout the school year. Members of the project staff constructed the practice tests for teachers to administer in their own classroom. The practice tests were designed to familiarize students with the procedures and formats of the standardized test used in their district. Additionally, the administration of practice tests provided students an opportunity to apply to a testing situation those test-taking skills taught in the filmstrips. The following sections will describe the rationale and procedures for the development of the practice tests.

Frequency. Originally, 12 practice tests were planned for administration to students in Experimental Groups I and II at an approximate rate of one test every two weeks. However, the construction of the practice tests became a much more complex task than had been anticipated, and the final number of practice tests produced was 7. A time line showing the production dates for the practice tests is included in Figure 3. Teachers administered the practice tests approximately every three weeks, from October through March.

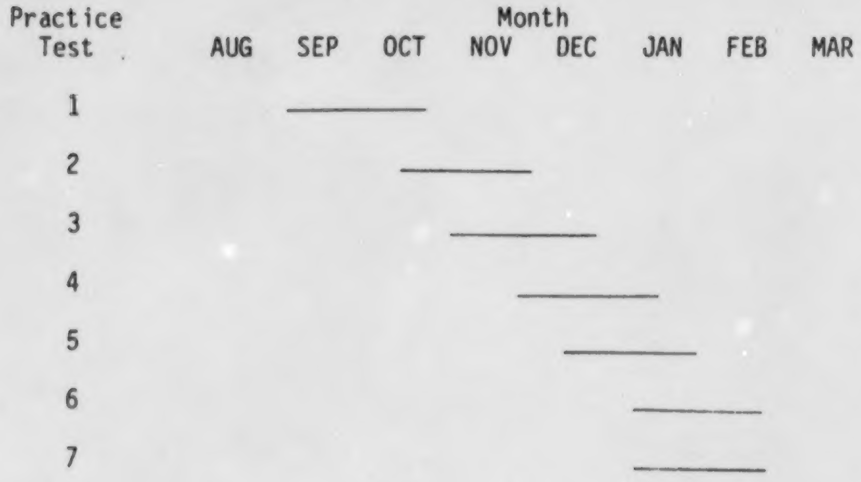


Figure 3. The time period for producing each practice test.

Practice tests were constructed to increase in length of time required for administration from 5 minutes (Test #1) to 30 minutes (Test #7). The gradual increase in time assisted the student in learning to work independently for the average number of minutes required to take one subtest on the actual test. The number of minutes and items for each practice test is displayed by school district in Appendix C. The mean number of minutes and items (across the four experimental districts) used for each practice test is shown in Table 15.

Table 15

The Mean Number of Items and Minutes Used for Each Practice Test

Practice Test	Mean Items	Mean Minutes
1	11.3	5.0
2	21.5	10.2
3	28.7	13.6
4	30.7	13.9
5	41.0	20.2
6	56.7	28.9
7	56.7	28.9

Format. Four different practice test series were developed. Each series was constructed to resemble the reading subtests (vocabulary, word analysis, and comprehension) used by the four districts participating in the study: Logan District (the pilot schools), the Metropolitan Achievement Test (MAT); Cache District, Comprehensive Tests of Basic Skills (CTBS); Granite District, the Stanford Achievement Test (SAT); and Nebo District, Iowa Tests of Basic Skills (ITBS). The other portions of the tests, such as science, math, social science, and language were not included in the practice test.

A copy of each of the standardized tests was obtained and the reading portion of the test was analyzed to determine how many items should be included in a 30-minute practice test. For instance, if the actual reading subtests required 90 minutes, only one-third the number of items (30/90) would be used for a 30-minute practice test. A chart showing this computation for practice test #7 (30 minutes) is located in Table 16. A proportional number of items was computed for the time limitation (5-30 minutes) for each of the seven practice tests. Each subtest within a practice test also contained a proportional number of items of those found in the actual test. Thus, if the actual standardized test was 56 minutes (see CTBS, 1973) and the vocabulary subtest was 15 minutes, a ratio of 15:56 would be maintained in the vocabulary subtest of the practice test. That is, vocabulary would be 8.1 minutes (15/56 X 30) and have 18 items (8.1 X 22). In this manner, each standardized test (MAT, SAT, CTBS, and ITBS) was examined and the appropriate number of items was computed for each practice subtest. (Copies of all practice tests are included in the Teacher's Manual.)

Another strategy employed in constructing the practice test format was to introduce only one or two reading subtests in each of the first several

Table 16

Computations Used to Determine Number of Items to Include in a 30-Minute Practice Test

TEST	SUBTEST	Items	Time	Items/ Minute	% Total Time	Proportion of 30 Minutes	Number[a] of Items
CTBS	Vocabulary	33	15	2.2	.27	8.1	18
1973	Sentences	23	20	1.15	.35	10.5	12
	Paragraphs	18	21	.86	.38	11.4	10
CTBS	Word Attack	40	38	1.05	.45	13.5	14
1981	Vocabulary	25	19	1.32	.22	6.6	9
	Comprehension	25	28	.89	.33	9.9	9
ITBS	Vocabulary- A	17	8	2.13	.12	3.6	8
	Vocabulary- B	13	6	2.17	.09	2.7	6
	Word Analysis	57	20	2.85	.29	8.7	25
	Pictures	23	12	1.92	.18	5.4	10
	Sentences	16	7	2.29	.10	3.0	7
	Stories	28	15	1.87	.22	6.6	12
MAT	Word Knowledge A	17	6	2.83	.10	3.0	9
	Word Knowledge B	23	12	1.92	.18	5.4	10
	Word Analysis	35	15	2.33	.24	7.2	17
	Reading A	13	7	1.86	.11	3.3	6
	Reading B	31	23	1.35	.37	11.1	15
SAT	Vocabulary	37	20	1.85	.22	6.6	12
	Reading A	45	20	2.25	.22	6.6	15
	Reading B	48	25	1.92	.28	8.4	16
	Word Study A	30	10	3.00	.11	3.3	10
	Word Study B	35	15	2.33	.17	5.1	12

[a] This number is the computed number of items for practice test #7. However, this number may be different from the number of items used in practice test #7 due to adjustments for standardized test formats (e.g., some subtests require items to be in groups of three).

practice tests until all subtests were included. For example, SAT Practice Test 1 was 4 minutes long and included only 7 Vocabulary items. In SAT Practice Test 2, Vocabulary was repeated (with new words) and a second subtest, Reading-Part A, was added. In Practice Test 3, Vocabulary was dropped and Reading-Part A, Reading-Part B, and Word Study Skills-Part A were included. Thus, one or two new subtests were progressively added to each practice test until all of the subtests from the actual reading test had been included. All of the reading subtests used in that particular standardized test were included in the last several practice tests.

Content. To generate the content for the practice test items, the actual reading series used in the four school districts were identified and texts obtained. Thus, the vocabulary words, comprehension skills, phonic sounds, and word attack skills in the practice test were those that students had actually studied in class. A complete list of the reading series used in the study is found in Appendix C.

Initially, teachers periodically informed the project staff about the pages they would be covering in their classes during upcoming weeks. Practice test items were then constructed using content from the reading series unit being taught at the time the practice test would be administered. For example, if the classroom was studying Unit 4 at the time the second practice test was administered, then the items drawn for practice test #2 would be from Unit 4.

The original plan was to construct three different practice tests for each classroom based on high, medium, and low reading levels found in most classes. Theoretically, it was possible that 120 different tests would be constructed for each of the seven practice tests because 40 teachers using three levels of different curriculum were participating in the study.

After identifying the content to be tested, items, correct answers, and distractors (wrong answers) were generated. To formulate distractors similar to those used in the actual standardized tests, the ITBS, CTBS, MAT, and SAT were closely examined and a list of the type of distractors used in the tests was constructed. These strategies are listed in Appendix C. For example, some of the construction strategies used in the standardized tests were words with similar final and initial sounds, words that were similar in appearance, and words with similar definitions or spellings.

After the project was started, it became clear that the production of 120 practice tests every two weeks (with anywhere from 10 to 70 items) was unrealistic. Based on the pilot testing of the first practice tests and considering the amount of time needed to generate practice tests and obtain feedback from the teachers on the pages covered in their reading tests, a more realistic procedure was developed.

Although the textbooks varied across teachers, the basic vocabulary, word attack, and comprehension skills were similar within reading level: high, medium, or low. Therefore, a generic list of vocabulary words and reading skills was generated for each practice test by surveying the texts within a reading level. The content for the four practice test formats was then drawn from the appropriate list and transformed into test items. This method resulted in students at similar reading levels receiving the same practice test content across districts but with a practice test format unique to their district.

Directions for practice test and scoring. Directions accompanying each standardized test were modified to fit the practice tests. Separate instructions were prepared for each subtest as the test items were prepared. (Complete copies of the directions for all practice tests are contained in the Teacher's Manual.) Different directions were written for students in Experimental Groups I and II. An example of the directions for one of the SAT practice tests (#5) for Experimental Group I is contained in Appendix C. Note that only one set of directions was necessary even though three levels of the practice test were administered in any given classroom at the same time. This could be done because even though most of the content for the three levels was different, sample items and any items in which the correct answer or stimulus was read verbally by the teacher (e.g., "Mark the word 'dog'") were the same.

Pilot testing. After the test items, distractors, instructions, and scoring keys were generated, the test in rough form was reviewed by project team members to detect major errors and inappropriate items. After necessary changes were made, blank formats and the draft test were sent to a graphics artist who drew the necessary pictures. Next, a typist inserted all the item content. Following the completion of the artwork and typing, the practice tests were reviewed again by staff for errors before the pilot test. Each level of each practice test was piloted with a small group of second grade children (two to three students per level of the test). The piloting was conducted to discover any typing errors, missing numbers or letters, and incorrect answer keys; to clarify instructions that were not easily understood; and to note misleading and ambiguous test items. Final adjustments were made, then the practice tests were mass produced and mailed to the teachers participating in the study.

Reinforcement Procedures

The Utah 79-80 State Refinements Project demonstrated that motivated students scored better on standardized achievement tests than students who were not motivated. However, this improvement in achievement test scores was attained by paying students money if they scored better than was predicted based on their pretest score. Clearly, it would not be practical to continue to pay students for trying hard on a standardized achievement test. For example, students would likely figure out that all they had to do was score poorly on the pretest to collect more money on the posttest. In addition, paying students based on their performance on a standardized achievement test would violate the norming procedures for the test. One of the goals of this project was to develop and evaluate a more practical alternative for motivating students to do their best on standardized achievement tests.

It was decided that violations of the norming procedures could be avoided by designing a motivational program to follow the biweekly practice tests. If students learned the habit of "trying their hardest" on the practice tests, hopefully, the habit would transfer and increase the students' motivation to try their hardest on the actual achievement test.

Rationale. To be effective, it was decided that the procedure developed should meet the following criteria.

1. Focus on effort, not aptitude.
2. Be motivating for the majority of students.
3. Remain motivating for the duration of the project (6 months).
4. Be minimally disruptive to the class that is using it and to the other classes in the school.
5. Require minimal time expenditure by teacher and students.
6. Require minimal monetary costs.

The use of tangible reinforcers such as a token economy did not meet several of the criteria listed above. For example, previous experience with token economies by the project staff indicated that although they are often initially effective, over long periods of time (as was the case with this project), token economies often lose their appeal and become difficult to maintain. Also, token economies are more of an exchange of goods or a payoff for performing well on a test instead of the desired intrinsic motivation to perform well on tests.

The strategy that best met the criteria stated above and was therefore selected for the student reinforcement component was a self-charting of improvement procedure. Self-charting of improvement refers to a procedure where students earn points which can be charted on a display (either public or private) for each increment of improvement on the targeted task. The effectiveness of this kind of a reward system to motivate students to perform academically has been demonstrated repeatedly (Paquin, 1978; Van Houten & Parsons, 1975; Willis, 1974). The self-charting materials and procedures are described below.

Description of procedure. Each Experimental Group I student received a personal chart mounted on a brightly colored poster board in a color selected by the student. The chart consisted of 7 horizontal bars, each bar representing 1 practice test. Each bar was divided into 50 segments, each segment representing one point (see Appendix D for a sample chart). Ample blank space remained on the chart and posterboard for the students to decorate the charts with their names and other creative artwork. The bar graph chart and the blank space for decorations allowed the students to personalize their charts freely in an attempt to make the charting process as individualized and reinforcing as possible. Each Experimental Group I classroom also received a 3 X 4 plywood display board equipped with 30 hooks on which the students' charts could hang. The teachers located the display boards in a prominent place in the room.

After the students scored their practice tests, they were to calculate the number of points their score exceeded an individually established criterion marked on their tests. This criterion was referred to as the student's "To Beat" score. Each point that equaled or exceeded the "To Beat" score was considered a "bonus point". The students were to graph the bonus points on their charts by marking the appropriate number of segments in the bar for that practice test. Approximately five minutes was given to graph and decorate the chart with crayons and colored pens (see Appendix D for a decorated chart). The charts were then returned to the display board for all to see.

Bonus points were cumulative. The points earned on each practice test were added to the points earned on subsequent tests for a new grand total. In this way, the students always had something to graph and decorate. When bonus points were earned, the students graphed the cumulative total (previous total plus bonus points). When no bonus points were earned, the previous total (plus 0) was entered on the chart. By allowing the bars to be graphed cumulatively, the charts always stayed the same height or grew taller. A decrease in points from one test to the next was never registered.

Additionally, because each child was given an individually established criterion to beat, the higher achieving students were not any more able to earn bonus points than the less able students. Thus, the reward system was set up so that students competed against themselves and other students to see how tall they could get their graphs to grow.

Project staff were responsible for determining the reinforcement criterion for each student. These "To Beat" scores were marked on each student's test before they were mailed to the teachers. Providing the student with the score that had to be beaten before taking the practice test was an attempt to increase the student's incentive to improve.

To determine the individual criteria for the first practice test, each teacher divided their classes into quartiles based on the information available at the beginning of the school year. Depending upon which quartiles they were in, the students were reinforced for scoring at or above the 20th, 40th, 60th, or 80th percentile of the test. On the subsequent practice tests, the students were reinforced for equaling or exceeding the percentage correct on the last test. For example, if a student's score was 15 on a 20-item test, i.e., 75% correct, the next test with 25 items would be assigned a "To Beat" score of 19 (also 75% correct). The average number of bonus points earned by students and the frequency with which students earned no bonus points can be used as an approximate indicator of whether the procedure was working. During the project, students earned an average number of 3.8 bonus points per practice test.

Pilot testing. Before implementing the reinforcement component in the Experimental Group I classrooms, a pilot test of the procedures was conducted in the four pilot classroom sites in Logan District. The procedures were observed by project staff and found to be executed as intended. Thus, the procedures were implemented as originally planned in the Experimental Group I classes.

Training Teachers in Standardized Test Administration

Although very little research has been done on the effects of quality of standardized test administration and student performance on the test, much has been done on factors which are related to the quality of test administration. As discussed in the review of related literature in Chapter II, factors such as rapport between the test administrator and students, anxiety on the part of students, whether students check their work, and the type of test instructions given are all related to students' performance on standardized achievement tests. The limited research which has been done underscores the importance of training teachers in standardized test administration techniques. For example, White, Taylor, Eldred, and Carcelli (1981) observed 38 teachers throughout Utah as a part of the 79-80 State Refinements contract and found that only 27% instructed students to check their work if they finished early, and less than 10% told students they should skip items that they do not know and go on to the next one. Even though teachers are instructed to do these things as a part of the standardized test Teacher's Manual, this previous State Refinements contract indicated that many teachers have difficulty following these instructions.

In another project, Taylor and White (1982) demonstrated that training teachers in test administration techniques substantially influences the scores received by students in those classes. Twenty-four classrooms were randomly assigned to an experimental group (classes in which teachers were specifically trained by the researchers in standardized test administration techniques) or control groups. Students in the 12 experimental classes scored approximately 1/2 standard deviation higher on the standardized achievement test than students who took the test from untrained teachers.

Materials utilized in the current project to train teachers in standardized test administration techniques were based on the materials from the Taylor and White (1982) project. Additions and refinements were made so that the training was more comprehensive and targeted more specifically on the standardized test being used by the participating districts. These materials were designed to provide skills to teachers in two areas: general standardized test administration techniques and administration techniques specific to the standardized test being used by each particular district. A brief description of the materials in each of these sections is provided below.

General standardized test administration procedures. General procedures for administering standardized achievement tests were presented and discussed in a workshop at the beginning of the school year. Topics covered during this workshop included the purpose of standardized achievement testing, pros and cons of groups versus individual testing, skills students need for standardized testing which are not generally required in other school work, how to motivate children, and a general review of what is required in a standardized administration (additional detail on these topics is included in the Implementation of Experimental Treatment section of this report and in the

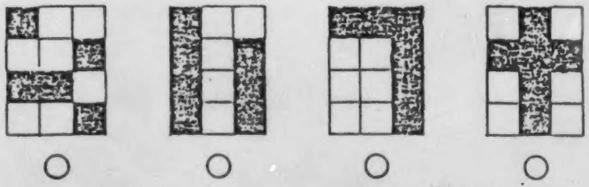
Presenter's Guide for Training Teachers in Test Administration which is available from the U.S. Department of Education).

Three primary types of activities were used during this workshop to stimulate discussion and present materials to the teachers. First, prior to the workshop, standardized achievement tests were analyzed, and items were selected to demonstrate to teachers the types of problems experienced by students on a standardized achievement tests. Because of these problems, the test results may be less valid for estimating what the student knows about a particular content area. Items or examples from standardized achievement tests were selected to demonstrate the following skills required during standardized achievement tests but which are not generally required during regular school activities.

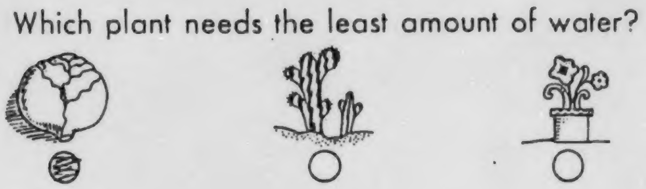
1. Selecting the "best" answer from a number of choices.
2. Eliminating attractive wrong choices.
3. Responding on machine-scorable forms.
4. Responding to specialized directions.
5. Working in a highly structured setting.
6. Responding with the whole class.
7. Identifying what question is being asked from a narrative.
8. Performing under time limits.
9. Following advice to guess.
10. Responding to unfamiliar figures or words.

For example, given below is one of the items taken from a standardized achievement test used to demonstrate to teachers how students sometimes have problems responding to unfamiliar figures or words.

10. "HERE ARE FOUR GARDENS, ALL THE SAME SIZE. THE DARK PARTS SHOW WHERE POTATOES HAVE BEEN PLANTED..." STUDENTS MAY THINK THEY ARE ON THE WRONG SET OF ANSWER CHOICES.



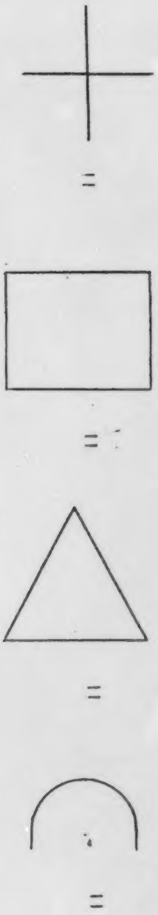
To demonstrate how children can sometimes know the correct answer to the content being tested but respond incorrectly on the test item, examples of problems were shown taken from the study by the Huron Institute where children were asked to explain why they had selected answers to questions. Given below is one of those examples.


Which plant needs the least amount of water?

When asked why she answered "cabbage", the child responded, "The cabbage needs the least water because it only needs water when you clean it." In other words, the child knew that since the other two options were growing plants, they would continually need water whereas the cabbage (which had been picked) would only need water when it was cleaned. Thus, the child knew the content but missed the item.

Items such as those presented above were used to demonstrate all of the areas of skills children need in responding appropriately to standardized achievement tests. This was done to help teachers understand the problems that students sometimes experience. It was hoped that such understanding would help teachers see the importance of structuring the testing situation in such a way that the student's knowledge of the content area is being tested rather than his or her skill in taking the standardized achievement test.

The second major activity used was a simulation activity. Teachers were asked to "take" a standardized achievement test consisting of one item. The directions for administering this test are given on the next page.



This is a test item to be administered to participants. Since this exercise provides experiences that illstrate points discussed in the workshop and stimulate much discussion, it is considered an important activity and should not be omitted.

Quickly read these directions to the test in a monotone and quiet voice while you keep your eyes on this paper. Move right into the test. Don't wait for participants to get oriented. Read the following exactly as written.

"Turn to HO/4. This is a hard test so listen up. Fill in the space above the correct answer to this question. You may not make any other marks on the paper. Which one of these would be the cheapest to buy? Listen carefully and I will tell you about what they cost. The hyperbola costs more than the triangle, the triangle costs the same as the plus, and the plus costs more than the square. Mark the cheapest one, the one that costs the least. Pencils down. Turn your papers over." (Correct answer: Square)

DO NOT TELL THE ANSWER UNTIL YOU HAVE GIVEN THE ITEM TWICE. Procede in this manner. Ask the first question below to generate discussion. Keep it brief and encourage one sentence responses. They will have much to say but try to get answers to questions 2 through 9 if they are not covered in the discussion.

1. *What did I do wrong?*
2. *What was the question?*
3. *How many times did I read the question? (2)*
4. *How did you place the paper in front of you?*
5. *What is a hyperbola?*
6. *Do you need to know what a hyperbola is to answer the question?*
7. *Did you stop listening after hearing "hyperbola"?*
8. *What strategy did you use to figure out the answer?*
9. *How many answered the question?*
10. *In this case, how much control does a test administrator have over the test results? (100%. Since virtually no one will get the item correct, but they do know the content, the examiner had control.)*
11. *This is a real test question, only the figures have been changed. The wording is otherwise untouched. What grade level do you think it is? (2nd)*

Readminister the item but follow correct test administration practice.

1. *Postive verbal reinforcement before and after testing.*
2. *Preparation of examinees (demonstrate how to turn paper upside down and fill in the answer forms).*
3. *Look at examinees at the end of each sentence to assure that they are responding correctly.*
4. *Pause at the end of each sentence.*

After reading the item, talk about anything the participants wish to discuss, but don't volunteer the answer. When someone finally asks for the answer, tell them that the rule is you can't tell them, marking the point that we don't tell students the answers. Then tell them the answer, ask how many got it right, and reward everyone for trying. Trhoughout the workshop, the presenter may use this test experience to illustrate other testing events that create problems for students.

After taking this test, participants were asked to discuss the experience that they had just had. This item generated a great deal of discussion about the proper and improper ways to administer standardized tests. Participants were particularly impressed to learn that the item is taken from a second grade test in which symbols instead of names of toys were used so that teachers might be unfamiliar with some of the lanquage (in the second grade test, a teddy bear, roller skate, football, and doll are used instead of plus, square, triangle, and hyperbola).

BEST COPY AVAILABLE

The third activity consisted of critiquing a videotape developed by the project. This tape was based on a videotape developed during the 79-80 State Refinements contract. The tape showed scenarios of standardized testing done both correctly and incorrectly. Although the scenarios on the videotape were acted out, all of the scenarios included on the videotape had actually been observed in the classroom. Teachers were asked to identify the correct and incorrect test administration procedures being done.

Administration Procedures Specific to a Particular Standardized Test

Shortly before the spring administration of the district's standardized achievement test, a second workshop was held in which teachers were provided additional training in administering the particular standardized achievement test being used by their district. The content of this workshop reviewed the general procedures for standardized test administration and then focused on the procedures for the particular test being used in that district. The review of general test administration procedures presented material taken from standardized test administration manuals and encouraged discussion from teachers based on their experience administering the practice tests during the year.

Even though all of these teachers had previously administered standardized achievement tests in their classroom, it was hoped that the combination of the workshop in the fall and the experience of administering a number of practice tests during the year would have sensitized them to a number of important issues about the administration of standardized achievement tests. For example, included in this discussion were issues such as student seating arrangement for testing, how to prepare for early finishers, clarifying ambiguities in the directions, and facilitating a supportive atmosphere for testing. A more detailed description of the types of materials covered in this workshop is included in the Implementation section of this report or in the Presenter's Guide which is available from the U.S. Department of Education.

Material presented in this workshop was developed based on the project staff's analysis of standardized test administration directions and their identification of areas which might cause some students to score lower on the test than would be accurate based on what they knew about the content area. The main learning activity consisted of the teachers in the group alternating in the role of the test administrator with portions of each subtest while the other teachers acted as "students". These roles were alternated so that each teacher had an opportunity to participate several times as a test administrator. Following each section, the group would discuss how the test was being administered, provide suggestions for improvement, and identify areas that might cause problems for students.

Summary

The purpose of this training in test administration was to sensitize teachers to the problems which students have during standardized test administration, to suggest the reason for many of those problems, emphasize how those problems might result in test scores being an inaccurate reflection of what the student knows about a particular content area, and to train teachers in techniques for substantially reducing or eliminating those problems. By focusing on examples from actual standardized achievement testing, simulated experiences for the teachers, and the videotape of test administration scenarios, an effort was made to make these points interesting and as "real life" as possible. The interactive nature of the training was an intentional part of the design, for although all of the teachers had previously given standardized achievement tests, almost none of them had done so in a situation where they could get feedback from others about their administration techniques.

Sample for Research

Potential participants in the research project were selected from three school districts located in central and northern Utah: Granite District is located in Salt Lake City, Nebo District in the south end of Utah County, and Cache District in Cache Valley (see Table 17 for a description of the districts). These districts provided an appropriate accessible population because they serve a large number and wide variety of Title I second graders and were accessible to the project base in Logan. Alpine District, originally proposed as a project site, was not included in the sample because an adequate number of teachers were available in the other three districts and the project logistics were simplified by working with three instead of four districts. The original sample contained 22 schools, 61 classes, and 1,448 students (see Table 18). One Cache Valley school, with two Experimental Group II classes, left the study in March due to unscheduled demands on the teachers' time; and one teacher in an Experimental Group II school in Granite District was dropped from the project in early February due to ill health in her family. This attrition resulted in a final sample of 21 schools with 58 teachers and 1,373 students. Experimental Group I had 21 classes and 522 students; Experimental Group II had 17 classes and 412 students; and the control group had 20 classes and 439 students. The process of determining the sample and the procedure for assignment to experimental groups is described below.

Identification and Selection of Sample

The process of selecting the participating districts began with an informational meeting which was held in May, 1981. District Title I coordinators from the Salt Lake and Utah Valley areas were invited to the meeting. Coordinators from 10 districts attended the meeting. Topics discussed during the meeting included previous research on standardized

Table 17

Description of Districts Participating in Project

District	Type of Geographic Area	Breakdown and Number of Schools	Enrollment	Ethnic Make-Up										Grades with Title I	Subject with Title I Services
				American Indian		Hispanic		Asian		Black		White			
				M	F	M	F	M	F	M	F	M	F		
Nebo	Rural	19 - Elementary	7,224	63	73	103	81	13	14	1	1	6341	6521	3-6	Reading only
		3 - Middle	1,798	.5%	.6%	.8%	.6%	.1%	.1%	-	-	48.1%	49.4%		
		3 - Junior High	1,744												
		4 - Senior High	2,417												
		29 - Total	13,193												
Granite	City	58 - Elementary	36,333	197	234	1198	1175	564	531	134	137	28824	27651	K-3 7-9 (1 school only)	Reading and Math
		14 - Junior High	13,118	.3%	.4%	2.0%	1.9%	.9%	.9%	.2%	.2%	45.9%	44.9%		
		8 - Senior High	12,376												
		80 - Total	62,827												
Cache	Rural	10 - Elementary	5,448	38	30	14	23	32	16	0	1	4325	4008	2-3	Reading Only
		2 - Junior High	1,794	.5%	.4%	.2%	.3%	.4%	.2%	0%	-	48.8%	45.2%		
		1 - Senior High	1,627												
		13 - Total	8,869												

Note: Taken from Annual Report of State Superintendent - USOE 1980-81, Utah Public School System.

Table 18

Experimental Sample

PILOT TESTING

District	School	Teacher
Logan District	Hillcrest	Larsen Peterson Olsen
Logan District	Riverside	Manley

EXPERIMENTAL GROUP I

District	School	Teacher
Granite District	Hillsdale	Jensen Kane Kunz Waldram
	Lincoln	Archer Norris
Granite District	West Kearns	Banks Borden Gomez Green Lobb Martin
	Redwood	Crockett Latham
Nebo District	Santaquin	Burbidge Payne
	Westside	Willis Anthony
Cache District	Wellsville	Jenkins Nielsen Murray

$\bar{X}$ 1981 Achievement Z Score

$\bar{X}$/SD = .11/.91

EXPERIMENTAL GROUP II

District	School	Teacher
Granite District	Western Hills	Cannon Eber Tanner Shepherd Schmidt
	Stansbury	Hunt Miller Wallace Archer
Granite District	South Kearns	Grose Madsen Franco
Nebo District	Goshen	Neff Boyack
	Wilson	Anderson Altenburg
Cache District	Lewiston	Mieure Schenevar
	Park	Taggart * Talbot *

$\bar{X}$ 1981 Achievement Z Score

$\bar{X}$/SD = .06/.98

CONTROL GROUP

District	School	Teacher
Granite District	Woodrow Wilson	Lund Cummings Jackson
	Roosevelt	Pugh Burton
Granite District	Lake Ridge	Belliston Woodland Spackman
Nebo District	Taylor	Smith Beaudin Ghiradelli
	Larsen	Jensen Lee
Nebo District	Brookside	Mason Lee
Cache District	Summit	Jensen Rawlins Mellvill
	Millville	Tuddenham Noble

$\bar{X}$ 1981 Achievement Z Score

$\bar{X}$/SD = .08/1.15

*Left project before completion.

testing, the results of the Utah 79-80 State Refinements Project, and a description of the proposed study. Reactions by all of the people at the meeting supported the value of a project such as this; and Granite, Nebo, and Alpine district coordinators said they would definitely like to participate in the project. A similar meeting was held the next week with the District Title I coordinators from Logan and Cache districts who also volunteered to participate in the project.

After the proposal was approved in July, 1981, the coordinators of Granite, Nebo, Cache, and Logan districts were contacted again, and procedures were initiated to obtain formal district approval for participation in the project. District coordinators were then supplied with a letter for them to revise as they wanted and send to the principals of the Title I schools in their district. The letter explained the project and requested that the principals encourage their second grade teachers to volunteer for the study. (A copy of the letter is included in Appendix E.) A list of the principals in Granite, Nebo, Cache, and Logan Districts to whom the letter had been sent was obtained from the district offices, and a project staff member contacted each principal by phone to determine if they would be willing to participate in the project. At this time, principals were informed that we did not yet know to which group (I, II, or control) their school would be assigned. This was done to avoid a threat to the internal validity of the study findings due to the experimental groups being volunteers. Because assignment to groups was not done until after it had been determined that all of the accessible population was willing to participate if selected, schools in the control group would be only randomly different from schools in the experimental groups on the variable of "volunteerism". Twenty-two of the twenty-three principals contacted agreed to participate contingent on the willingness of the individual second grade teachers. The only principal who declined said he would like to participate if he could be guaranteed a slot in the Experimental I group. Because this would have compromised the integrity of the experimental design as described above, his school was dropped from consideration.

A list of the second grade teachers from the interested schools was obtained during the second and third weeks in August, 1981. Project staff contacted each teacher by phone to explain the purpose of the study, the procedure for random assignment of classes to experimental groups, and the responsibilities the teachers would have if they were selected for the project. Again, teachers were not told in which group they would actually be since assignment to groups was not done until a sufficient number of teachers had volunteered for all three groups. Responsibilities of treatment group teachers included showing biweekly filmstrips, giving practice tests to their students, and attending one or two workshops. An honorarium of $25 (for Experimental Group II) and $50 (for Experimental Group I) was given to teachers for participating. Teachers in treatment and control groups were told that observers would collect data during the spring administration of the standardized achievement test. Sixty-one teachers out of the 66 contacted volunteered to participate in the study. The reasons for unwillingness to participate were a lack of willingness to risk being assigned to the control group, previous time commitments, or health problems in the family.

Assignment of the Sample to Groups

Schools instead of classes (i.e., teachers) were randomly assigned to one of the experimental or control groups. This assignment method ensured that all teachers in the same school were using the same treatment procedures and reduced the possibility that the treatment implementation would be contaminated by conversations and sharing of materials by teachers in the same school but using different "treatments". It was expected that teachers from different schools were less likely to share information and materials than teachers from the same schools. To assist in assigning schools to one of the three experimental groups, the previous spring's average achievement test score for the second or third grade of each school was obtained. Because the districts use different achievement tests, each school's score was converted into standard Z scores (within each district) so that each score was on a roughly comparable metric. The names of the participating schools were then randomly drawn from a box and assigned to either Experimental Group I, Experimental Group II, or the control group. After assignment, the average achievement Z score for each group was calculated to determine if the randomization procedure had resulted in approximately equivalent groups, which it had not. The random assignment procedure was repeated once more at which time equivalent groups in terms of previous year's achievement test Z scores were obtained (average Z scores and number of classes for each group are shown in Table 18).

During the last week in August and the first week in September, each teacher was phoned and informed of the group to which they had been assigned and the specific responsibilities they could expect while participating in the project. A follow-up letter was sent to each teacher confirming their participation in the project and their group assignment (see Appendix E for a copy of a letter).

Implementation of Experimental Treatments

The following sections discuss the procedures for implementing the research treatments designated Experimental I or Experimental II (see Table 19 for the number of classes and students receiving each treatment). Experimental I classrooms received teacher training in test administration and student training in test-taking skills (including filmstrips, student practice tests, and student reinforcement for practice test performance). Experimental I classrooms received the filmstrips and the student practice tests. No experimental treatments were applied to control group classrooms. The four treatments are described below in two sections: teacher training in test administration and student training in test-taking skills (filmstrips, practice tests, and reinforcement). Table 20 displays the implementation time line for all components.

Table 19

Implementation of Experimental Treatments

	N		Teacher Training	Student Training		
Group	Classes	Students	Test Administration	Filmstrips	Practice Tests	Reinforcement
Experimental I	21	522	X	X	X	X
Experimental II	17	412		X	X	
Control	20	439	-	-	-	-

Teacher Training in Test Administration

The Utah 79-80 State Refinements Project (1981) concluded that the procedures used by teachers during test administration contribute to how well a student scores on a test. The data from that project also provided evidence

Table 20

Actual Timeline for Implementing Filmstrips,
Practice Tests, and Teacher Supervision

FILMSTRIPS	SEP				OCT				NOV				DEC				JAN				FEB				MAR				APR
#1				X...	....	..																							
#2								X.	...																				
#3												X	...	...															
#4																	X.	...	...	...	...	..							
#5																					X	...	...	..					
#6																						X	..	...	...	..			
#7																									X	...	...		
#8																										X	...	..	
#9																											X	...	
PRACTICE TESTS																													
#1								X ..	...	...	...																		
#2												X		...	...	...	..												
#3																		X ...	...	...	...	..							
#4																				X	....	..							
#5																							X	...	...	...			
#6																									X	...	...	.	
#7																												X...	
TEACHER SUPERVISION																													
Train Experimental I		X																											
Train Experimental II			X																										
On-site Model				...	...	...	..																						
On-site Visits								...	...								...	...	.										
Phone Visits													...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
Group Meeting																			..	...									

Note. X = deliver or mail materials.
. = implementation.

that teachers trained in proper standardized test administration had much higher levels of on-task behavior and quality of test administration than did untrained teachers, and students in the classrooms with trained teachers made significantly fewer errors in completing their test booklets.

This project, building on the results from the previous project, developed, implemented, and evaluated the effectiveness of a more extensive and pragmatic program designed to increase the quality of test administration. The program incorporated not only general test administration techniques but procedures specific to the actual standardized achievement test used by each district as well.

Only those 21 teachers assigned to Experimental Group I participated in the program for training teachers in standardized test administration. This training was presented in two structured workshops: the fall workshop was conducted in September at the beginning of the project, and the spring workshop was prior to the districts' spring achievement testing (see Table 21 for a breakdown by district).

Table 21

Training in Test Administration
Breakdown by District

Workshop	District	N	Date	Duration
Fall	Cache	3	September 12, 1981	2 hours
	Granite	11	September 12, 1981	2 hours
	Nebo	4	September 12, 1981	2 hours
(Make-up)	Granite	3	September 19, 1981	2 hours
Spring	Cache	3	March 11, 1982	3 hours
	Granite	14	March 12, 1982	3 hours
	Nebo	4	March 11, 1982	3 hours

The goals, agenda, implementation procedures, and teacher evaluations of each workshop are described below.

Fall workshop. The fall workshop was presented by five project staff in Salt Lake City on September 12, 1982, from 9:00 a.m. to 4:00 p.m. The three absentees, all from Granite District, participated in a make-up workshop on September 19, 1982 (see Table 21).

The primary purpose of this workshop was to train Experimental Group I teachers in the general procedures of proper standardized test administration. It was conducted in the fall for two reasons. First, techniques presented enabled the teachers to practice proper test administration procedures while administering the seven student practice tests described earlier. Secondly, other project-related information concerning the student training materials, the purpose of the research, and other logistical information needed to be given to teachers at the beginning of the project. Because this workshop was already scheduled, it was a natural time to include the training in test administration as one part of the workshop.

The workshop objectives which pertained to training teachers to administer standardized tests were as follows:

Participants will be able to:

1. Identify testing problems unique to the school district.
2. Differentiate behaviors required of teachers and students during testing from behaviors exhibited during the regular instruction.
3. List motivational, test-taking, and test administration practices that increase the validity of test results.
4. Produce
    a. a list of potential test-taking reinforcements.
    b. a statement of testing purpose for explaining to students.

5. Practice

   a. taking a test.
   b. teaching test-specific directions.
   c. completing a checklist of appropriate test administration practices.
   d. using the Teacher Index to Valid Test Performance.

6. Identify correct and incorrect test administration practices in videotaped classroom testing scenes.

A brief summary of each agenda topic is provided below (for more complete information, see Presenter's Guide for Training Teachers in Test Administration).

I. Introduction: Participants identified testing problems, took a simulated test, and filled out the Participant Inventory so they could assess their own pre-workshop knowledge of proper test administration (see Presenter's Guide).

II. Valid Test Results: The goals of achievement testing and the concept of validity were discussed. Factors that contribute to low test scores were presented as well as the advantages and disadvantages of group testing.

III. Motivation: Techniques to structure the environment to encourage students to try their best were presented and discussed.

IV. Test-Taking Skills: Student skills required during test taking but which are not generally required during regular school activities were explained and simulated with actual achievement test items.

V. Test Administration: Techniques for obtaining more valid results were presented. The teachers practiced these procedures while administering sections of a standardized achievement test to each other. The Quality of Test Administration Checklist (located in the Presenter's Guide) was presented and discussed. The Teacher Index to Valid Test Performance form (located in the Presenter's Guide) to document disruptive events that may occur during testing was explained.

VI. Videotape Observation: A videotape developed during the Utah 79-80 State Refinements Project (1981) was shown to illustrate the effect of various test administration procedures on student behavior. Scenarios depicting both correct and incorrect test administration techniques were critiqued by the participating teachers. The following testing activities were shown in the videotape: preparing students for the test, arranging the testing room, distributing the test materials, giving directions, monitoring students, using an aid, providing assistance to the students, pacing, and obtaining group responses.

VII. Summary: The teachers took the Participant Inventory again so they could assess the degree to which they had acquired the skills and information presented to them during the workshop.

VIII. Feedback and Written Evaluation: An evaluation form was distributed to all the teachers and collected at the end of the workshop. Results, shown in Table 22, indicate that the workshop was very successful in meeting the objectives of the project and the perceived needs of the participants.

Spring workshop. Three spring workshops, one in each participating district, were presented on March 11 or 12, 1982 (see Table 21). Separate workshops for each district were held by project staff in the schools to enable the teachers to attend right after school. The workshop was conducted in the spring to increase the likelihood that the information provided would be recalled and used by the teachers when actually administering the districts' standardized achievement tests in April. Each workshop was approximately 3 hours long. There were no absentees during those workshops.

The primary purpose of the spring workshop was to train Experimental Group I teachers in the specific test administration procedures relevant to the district-adopted achievement test they would be administering to their students in April.

The workshop objectives were as follows:

Participants will be able to:

1. Administer the publisher's practice test using proper test administration techniques.
2. Administer the standardized achievement test to their students with proper test administration.

Items from the spring workshop agenda are summarized below. (A copy of the spring workshop materials is included in the Presenter's Guide.)

I. Things to Do: This topic included specific activities for the teacher to do before the testing date, just before testing, during testing, and after testing.

Table 22

Fall Workshop Evaluation Data

Workshop Evaluation Form

September 12, 1981 — Date
Salt Lake City — Location
N = 19

I. EVALUATION OF WORKSHOP STAFF

KNOWLEDGE OF SUBJECT MATTER

19 Very well informed
___ Adequately informed
___ Not well informed
___ Very poorly informed

ATTITUDE TOWARD SUBJECT

19 Enthusiastic
___ Rather interested
___ Routine interest
___ Disinterested

ABILITY TO EXPLAIN

17 Clear and to the point
2 Usually adequate
___ Somewhat inadequate
___ Totally inadequate

LEVEL OF PRESENTATION

15 Very well suited to participants
4 Moderately well suited to participants
___ Completely above participants
___ Completely below participants

ATTITUDE TOWARD PARTICIPANTS

16 Very helpful and understanding
2 Interested
1 Routine, neutral
___ Distant, cold, aloof

METHOD OF PRESENTATION

6 Ingenious, creative
13 Interesting, held attention
___ Somewhat monotonous
___ Uninteresting, boring

OPPORTUNITY FOR DISCUSSION

___ Too infrequent
18 Appropriate
1 Too frequent

OVERALL RATING OF WORKSHOP STAFF

14 Outstanding
5 Better than average
___ Average
___ Below average
___ Poor

II. EVALUATION OF WORKSHOP CONTENT AND FORMAT

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree	
	Frequency					$\bar{X}$
• The objectives of the workshop were clear from the beginning.	0	0	1	9	9	4.42
• The balance between lecture and participant interaction in the workshop was ideal	0	0	0	16	3	4.16
• The workshop material contributed well to our overall goals and objectives.	0	0	0	8	11	4.58
• The workshop was well structured and organized.	0	0	0	7	12	4.63
• The content of the workshop was presented in a clear and understandable manner.	0	0	0	9	10	4.53
• The scope and coverage of this workshop was appropriate	0	0	2	9	8	4.32
• Content was summarized well and major points were easy to identify.	0	0	0	13	6	4.32
• The value I derived from this workshop was well worth the time required of me to participate	0	0	1	10	8	4.37
• The workshop provided specific guidance and ideas which I can apply in my job responsibilities.	0	0	0	8	11	4.58
• The total length of the workshop was appropriate.	0	3	2	14	0	3.58
• Workshop arrangements (location, rooms, prior information, schedules) were adequate	0	2	2	14	1	3.73

BEST COPY AVAILABLE

Table 22 (continued)

III. OVERALL EVALUATION

1. OVERALL RATING OF WORKSHOP

13 Outstanding

6 Better than average

___ Average

___ Below Average

___ Poor

2. Specific points which were valuable or significant to me were: (list at least two)

Reinforcement/motivation	10
Videotape on test administration	1
Practice tests/group response	4
Introduction to test-taking/ examples of difficult items	9
Good visual aids	1
Other uses for test skills	1
Filmstrip	1
Role playing of students	1
Good workshop staff	1

3. The workshop would have been more valuable to me if: (list at least two, particularly refer back to items you rated low in first two sections)

Split to 2½ days	2
If I'd had a choice about participating	1
Too warm	1
Closer with less travel	1
Practice test was too long	4
Shorter lunch	2
Shorter workshop	2
Listing do's and don'ts on videotape	1
Nothing	3

[illegible]. If you had to shorten this workshop by ½ hour, what would you delete?

Nothing	2
Going through practice test	4
Generally condense	2
Practice direction giving	7
Gotten lunch orders at first	1
Percentages about student and teacher performance	1
1st group question-answer period	1
2½ day sessions	1

II. Understanding the Nature and Purpose of the Test: The purpose of standardized testing, assumptions of the publisher, and type of subtests in each test was presented and discussed.

III. Schedule: Strategies for timing, breaks, avoiding testing days close to holidays or special events, and use of the school day were presented.

IV. Use of Proctor or Aide: Proctor/student ratios, classroom management, and test management with the use of a proctor or aide was presented.

V. Informing Students and Parents of Impending Test: Procedures for informing students and parents about the testing schedule, what will be tested, how the results will be used, special preparation for the student, and student concerns were discussed.

VI. Seating: The use of separate desks, proper desk positioning, and teacher contact during the test was encouraged.

VII. Early Finishers: Teachers were instructed to remind students to check their work and provide a nondisruptive task, such as drawing, for early finishers.

VIII. Eliminating Distractors: Teachers were warned of potential distractors with suggestions on how to minimize them.

IX. Facilitating a Supportive Atmosphere: Student anxiety about test-taking was discussed with suggestions on how to create a supportive atmosphere.

X. Reading Directions Carefully/Clarify Ambiguities: Proper procedures for reading directions were outlined. Teachers were informed of the extent to which they may add directions for the purpose of clarification, or otherwise assisting the students.

XI. Monitoring Students: Unobtrusive and supportive ways of monitoring students were presented to prevent cheating, discourage random guessing, and prompt dawdlers.

XII. Answering Student Questions: Teachers were informed of the benefits of responding to student questions about specific test items after the test is completed. Suggestions about managing such a classroom discussion were provided.

XIII. Preparation of the Test Booklet for Scoring: Teachers were instructed about responsibilities such as erasing extraneous marks on student booklets, darkening circles that were filled in too lightly, copying over tests that were ripped, and situations which may necessitate invalidating a subtest.

XIV. Use of Valid Test Performance Index: A rationale for the use of the Valid Test Performance Index to document disruptive events during testing were provided. Teachers were instructed on how to use the Index.

XV. Practice/Review of Standardized Achievement Tests: Each teacher was provided with an analysis by subset of the test they would be using. The analysis included a description of each subtest, test vocabulary, and time limits, and notes for giving directions keyed to specific items on the test (see Presenter's Manual, Spring Workshop Materials). They practiced administering selected items from each subtest in role play situations.

XVI. Practice/Review of Publisher's Practice Test: The rationale for using the publisher's practice test and strategies for using it to its optimal benefit were provided. The teacher practiced administering the practice test in role play situations. (See Spring Workshop Materials, Presenter's Guide).

XVII. Feedback and Written Evaluation: Written documentation of the teacher's evaluation of the workshop was obtained on the Final Project Evaluation Form (see Table , items 36-39).

Student Training in Test-Taking Skills

This section discusses the implementation of the three student training components described earlier: filmstrips, practice tests, and reinforcement procedures. Since the implementation of the three components is so interrelated, activities are presented chronologically and refer to both Experimental Groups I and II classrooms except for the reinforcement procedures or where otherwise indicated.

Training teachers to implement student training components. To train Experimental Group I and Experimental Group II teachers to implement the student training components, two workshops were conducted in fall, 1981, by five project staff. Eighteen Experimental Group I teachers were trained to implement the Treatment I components (filmstrips, practice tests, and reinforcement procedures) in conjunction with the Test Administration Workshop on September 12, 1981. Twenty Experimental Group II teachers were trained to use the filmstrips and practice tests at a workshop held in Salt Lake City on September 19, 1981. Each workshop was four hours long. Three Experimental Group I teachers from Granite District who could not attend the workshop on September 12 were trained on September 19 with the Experimental II teachers (see Table 23 for a breakdown).

Table 23

Workshop in Student Training Implementation
Breakdown by District

Experimental Group	District	N*	Date	Duration
I	Cache	3	September 12, 1981	4 hours
	Granite	11	September 12, 1981	4 hours
	Nebo	4	September 12, 1981	4 hours
(Make-up)	Granite	3	September 19, 1981	4 hours
II	Cache	4	September 19, 1981	4 hours
	Granite	12	September 19, 1981	4 hours
	Nebo	4	September 19, 1981	4 hours

*Note. Three Experimental Group II teachers left the project before completion (two from Cache and one from Granite).

There were three goals for the Student Training Materials Implementation Workshop:

1. To train teachers in the use of the student training components to which they had been assigned.

2. To train teachers in the documentation and communication procedures necessary for project operation.

3. To schedule the student training dates and collect the curriculum information necessary to develop the practice tests.

A brief summary of each agenda topic is presented below.

I. Overview: The findings of the Utah 79-80 State Refinements Project, the research objectives and outcome measures for this study, and a brief introduction to the treatment components were presented.

II. Basic Instructional Philosophy and Procedures: The rationale for using a direct instructional approach was discussed and the procedures (model, lead, test, and correct) were explained.

III. Plan for Student Training: The schedule for implementing the student training components throughout the year was presented.

IV. Filmstrip Training Package: The interactive format of the filmstrips, topics covered in each filmstrip, and workbook activities were explained. Segments of Filmstrips #1 and #2 were shown to the teachers as they played the role of second grade students. This illustrated the procedures necessary for the proper implementation of the filmstrip package.

V. Practice Tests: The rationale for training students in test format and the procedures for administering the practice tests were explained. The teachers role-played second grade students as they took a sample practice test. This illustrated the proper administration procedures for the use of the practice test component.

VI. Reinforcement Procedures (Experimental Group I teachers only): Teachers were presented with the rationale for using the reinforcement procedures. Implementation of the procedures including scoring of the practice tests, training the students to calculate the bonus points they earned, and using the reinforcement chart were explained. The teacher went through the procedure as they role-played second grade students.

VII. Communication Procedures: The procedures for returning the biweekly tests, updating project staff about reading curriculum progress, maintaining accurate records of attendance on the appropriate form, phone consultations, and on-site visits were explained.

VIII. Yearly Scheduling: Teachers and project staff scheduled their first filmstrip and on-site visit and outlined the expected curriculum progress for the year. Contact logs to document the communication between the teachers and project staff were presented.

IX: Feedback and Written Evaluation: An evaluation form was distributed and completed by all the teachers. Since the September 12, 1981 Student Training Materials Implementation Workshop was held for Experimental Group I teachers concurrently with the Test Administration Workshop, the results of both workshops were simultaneously on the same form and are reported previously in Table 22. The results of the September 19, 1982 workshop are presented in Table 24. Findings indicate that both workshops successfully met the goals.

Teacher's Manual. In addition to the workshop training, a Teacher's Manual was developed to provide the participating districts with all the materials needed to implement the student training (with the exception of filmstrips and tapes which were included in a separate package). The manual, How to Take Tests--Team Teaching with Professor Owl, includes all the written student training curricula produced for the project and the rationale supporting the format and content used. It is arranged in three sections: Filmstrips, Practice Tests, and Reinforcement and provides instructions for using the material, master copies of consumable items, and supplementary activities for review.

Table 24

Teacher Training in Student Curriculum Workshop

Workshop Evaluation Form

September 19, 1981
Date

State Office of Education
Location

N = 18

I. EVALUATION OF WORKSHOP STAFF

KNOWLEDGE OF SUBJECT MATTER

12 Very well informed
6 Adequately informed
___ Not well informed
___ Very poorly informed

ATTITUDE TOWARD SUBJECT

16 Enthusiastic
2 Rather interested
___ Routine interest
___ Disinterested

ABILITY TO EXPLAIN

11 Clear and to the point
7 Usually adequate
___ Somewhat inadequate
___ Totally inadequate

LEVEL OF PRESENTATION

15 Very well suited to participants
3 Moderately well suited to participants
___ Completely above participants
___ Completely below participants

ATTITUDE TOWARD PARTICIPANTS

15 Very helpful and understanding
3 Interested
___ Routine, neutral
___ Distant, cold, aloof

METHOD OF PRESENTATION

6 Ingenious, creative
12 Interesting, held attention
___ Somewhat monotonous
___ Uninteresting, boring

OPPORTUNITY FOR DISCUSSION

___ Too infrequent
17 Appropriate
1 Too frequent

OVERALL RATING OF WORKSHOP STAFF

10 Outstanding
8 Better than average
___ Average
___ Below average
___ Poor

II. EVALUATION OF WORKSHOP CONTENT AND FORMAT

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree	
	Frequency					$\bar{X}$
• The objectives of the workshop were clear from the beginning.	0	0	0	12	6	4.33
• The balance between lecture and participant interaction in the workshop was ideal	0	0	1	11	6	4.27
• The workshop material contributed well to our overall goals and objectives.	0	0	1	9	7	4.35
• The workshop was well structured and organized.	0	0	1	10	7	4.33
• The content of the workshop was presented in a clear and understandable manner.	0	0	1	11	6	4.27
• The scope and coverage of this workshop was appropriate	0	0	0	11	7	4.38
• Content was summarized well and major points were easy to identify.	0	0	0	12	5	4.29
• The value I derived from this workshop was well worth the time required of me to participate	0	0	5	7	6	4.05
• The workshop provided specific guidance and ideas which I can apply in my job responsibilities.	0	0	4	9	5	4.05
• The total length of the workshop was appropriate.	0	0	3	10	5	4.11
• Workshop arrangements (location, rooms, prior information, schedules) were adequate	0	0	0	16	2	4.11

BEST COPY AVAILABLE

III. OVERALL EVALUATION

1. OVERALL RATING OF WORKSHOP

_8_ Outstanding

_10_ Better than average

___ Average

___ Below Average

___ Poor

2. Specific points which were valuable or significant to me were: (list at least two)

Simulated test item	1
Filmstrip presentation	5
Sample test	1
Demonstrate with children (so they see it done)	
All materials prepared for teacher	
Standardized testing is not a part of teacher training	3
Statistics (of test formats)	3
Role of teacher in testing	1
Answered questions of teacher role in project	3
Help students	

3. The workshop would have been more valuable to me if: (list at least two, particularly refer back to items you rated low in first two sections)

More input on filmstrip and prep.	1
Coffee	
Hard to give up a Saturday. Not sure of work involved in project.	3
More ?'s on group standardized testing.	
Implement concepts.	
Baby was distracting.	

The manual was developed and added to as the student training components were produced. The 21 Experimental Group I and 17 Experimental Group II teachers received a large three-ring notebook with labeled dividers to bind and organize the material which was sent to them with each filmstrip and practice test. The materials were hole-punched and ready for inserting into the manual. A listing of the materials in each section is provided below.

I. Introduction: Organization of the manual and materials.

II. Filmstrips:

General Information--rationale for student training.
Teacher/Filmstrip--interaction of filmstrip instruction with teacher behavior.
Instructional Sequence--explanation of direct instruction strategy.
General Instruction--tasks required to show filmstrips.
9 Filmstrip Scripts--for teacher preparation and for the projectionist to use in turning the frames.
9 Masters for Work Booklets--for duplicating student practice material.
Practice and Review Sheets--laminated or master copies for supplementary activities to use as needed or just for fun.

III. Practice Test Section:

General Information--a rationale for training students in test format.
Construction of the Practice Tests--explanation of the development of the tests.
Procedures--how to use tests properly.
Directions--explanation of the individual test directions.
General Procedures--instructions to the teacher for administering the practice tests.
Scoring Procedures--directions for instructing the students how to score the practice tests.
Laminated Scoring Cards--for giving students examples of scoring.
7 Practice Test Masters--to duplicate for distribution to students.
7 Practice Test Directions--individual test directions to direct students through each test.

IV. Reinforcement Section:

Motivation Program--explanation of the rationale for the program and procedures for implementation.
Charting Instructions--specific instructions to teach children to use program.
Sample Charts--for explanation to students.
Master Chart--for reproduction as needed.
Laminated Chart--for teaching students how to fill in graphs and calculate points.

Typical implementation. The typical cycle used to implement student training proceeded through teacher preparation, showing the filmstrip, practice test administration and scoring, reinforcement implementation, and return of tests to USU. The details of each activity are explained below.

Upon receiving the classroom materials from USU, the teacher would prepare the lessons and schedule related activities to occur within two weeks. The average time for teacher preparation for conducting a typical filmstrip and practice test (according to self-report information) was 11.9 minutes and included the following:

1. Duplicate extra practice tests and filmstrip booklets from the master copy for any new students.
2. Arrange for someone to turn the filmstrip projector.
3. Read script accompanying filmstrip (optional).
4. Read test directions (optional).
5. Post or copy for each student the review charts.
6. Position filmstrip projector and tape recorder in room.

To implement the filmstrip and tape lesson, the teacher would first use review charts to prompt students on concepts taught in previous filmstrips. (See the review charts that accompanied each filmstrip lesson in the Teacher's Manual.) After a short review of 2-5 minutes, the teacher would pass out individual student booklets and start the tape while the filmstrip turner ran the projector. Most of the instruction was delivered to students via the filmstrip, but the teacher could control the pace to the degree she/he wanted to and would personally direct the class for three types of exercises:

1. when asked by Professor Owl to teach or quiz the students on a difficult concept,
2. to supplement the filmstrip instruction when students were having difficulty understanding,
3. to supervise students as they worked through practice items in their booklets.

Following the filmstrip, students could take their booklets home. The practice test was given on a different day from the filmstrips. Although there was not a one-to-one correspondence between the objectives of a particular filmstrip and the following practice test, the tests were usually scheduled in between two filmstrips. Typically, the practice test would be given one or two days following a filmstrip.

Before administering the practice test, the teacher was encouraged to review previously taught test-taking skills. After passing out the individually identified tests to the correct students, the teacher would begin reading directions and giving the test. The directions were structured so that all three levels could be given at the same time, yet students would not realize that tests differed depending on the reading level. The length of the test and the time allowed for completion increased with each practice test (5 minutes on test #1 to 30 minutes on test #7).

Immediately following the test, red pencils, supplied by the project, were passed out to the students. Black lead pencils were put inside desks. Scoring directions were reviewed according to the needs of the students. As the teacher read the answers, students marked their own papers. Then the number of correct answers were tallied and placed in a box marked "score" on the test cover. The mean percent correct for all practice tests was 82.75 (SD = 14.72).

In addition to viewing filmstrips and taking practice tests, Experimental Group I students participated in the reinforcement procedures. They were verbally encouraged to try their best to score high on the practice tests and beat a cut-off score assigned to them based on a previous test score. Reinforcement procedures were initiated immediately after the students scored their tests. The front cover of the individual practice tests contained three labeled boxes (see the Teacher's Manual for the cover sheets): "SCORE", "TO

BEAT", and "POINTS". Each test contained an individually predetermined "TO BEAT" number based on his/her score on the previous practice test. After filling in their "SCORE" boxes, students would compute their "POINTS" by subtracting "TO BEAT" from "SCORE". Assistance in computation was given by the teacher for students who could not subtract. The mean number of reinforcement points earned per student per practice test was 3.8 (SD = 1.4).

Students obtained their reinforcement charts from the reinforcement stand and copied the data from test to chart. Charts were graphed and shown to the teacher, who was encouraged to praise the students for progress. Students then returned the chart to the stand for public display and gave their tests and red pencils to the teacher.

Teachers were instructed to record the names of students who were absent during the filmstrips and practice tests. When convenient, absent students "made up" the test and filmstrip. Tests were then mailed back to the project for analysis and filmstrips were either passed on to another teacher or stored for later use.

Supervision. Throughout the project year, contact with teachers was maintained by USU staff through classroom and phone visits. Table 25 displays the frequency and types of USU-teacher contact. The interactions with teachers served two purposes:

1. To support and reinforce the teachers during their facilitation of project components. A higher degree and quality of implementation was expected from teachers who were contacted and rewarded frequently.
2. To correct problems and modify implementation strategies to fit unique situations. The most efficacious method to ensure proper implementation of a program is to stop misconceptions at the

inception, model the correct procedure, and maintain frequent follow-up.

Table 25

Number of Contacts Between Project Staff
and District Staff

District	Number of Teachers	Model Procedures			Observe Procedures			Number of Phone Calls
		Filmstrips	Practice Tests	Reinforcement	Filmstrips	Tests	Reinforcement	
Nebo	8	15	13	8	17	11	8	40
Granite	25	16	22	16	32	15	14	236
Cache	5	5	5	5	5	-	-	25

As indicated earlier in Table 20, supervision of teacher implementation by project staff began soon after teachers were trained to use the components. In conjunction with the first two filmstrips and the first practice test, staff visited all classrooms to model procedures and observe the components being used. The average number of visits made per teacher was 4.1. After the initial visits, follow-up observations were made to those teachers who needed more assistance.

Teachers who were judged to be implementing the program correctly were phoned periodically to discuss progress. As indicated in Table 20, phone visits were conducted from December 1, 1981 to April 23, 1982. An average of 7.9 phone visits per teacher were made.

From January 18 to February 1, staff members conducted small group meetings with teachers by school. During this time, teachers were asked to express their positive and negative feelings toward the project. Ideas for more efficient implementation and management were shared. Project staff suggested methods for smoother operations. Feedback from teachers concerning filmstrips and practice tests was recorded and recommendations about how the project could be improved were noted.

Delivery of materials. Project materials (filmstrips and practice tests) were periodically hand carried or mailed to teachers (Table 20 indicates the delivery dates). Teachers scheduled the filmstrips on different dates from the practice tests and within two weeks of receiving materials. Individual student reinforcement boards and classroom reinforcement stands were delivered to each teacher in Experimental Group I prior to the first practice test administration.

Absentees and attrition. Students who were absent the day a practice test was administered or a filmstrip shown were given make-ups whenever possible. Data were kept by the teacher to show which students did not participate in which activities so that absenteeism could be accounted for in the data analyses. Tables 26 and 27 show the number of students who were absent and present for each filmstrip and practice test. The mean class attendance for filmstrips and for practice tests was 25. The mean class absenteeism was .9 student per filmstrip and 1.4 students per practice test.

Evaluation of project implementation. Data used to evaluate the project implementation came from three sources: teacher judgments about individual filmstrips, practice tests, and reinforcement procedures; teacher judgments about the project as a whole; and staff judgment on the quality of individual teacher implementation.

Filmstrips. Evaluation data on the filmstrips were collected on filmstrip evaluation forms (see Appendix F) which the teacher filled in and mailed to USU immediately after showing each filmstrip. No evaluations were conducted on Filmstrips #1 and #2 because staff were in

Table 26

Number of Classrooms, Students Present, and Students Absent for Each Filmstrip

District	Experi-mental	FILMSTRIP																									X̄		
		1			2			3			4			5			6			7			8			9			
		Classes	Students Present[a]	Students Absent[b]	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent
Cache	I	3	83	0	3	82	1	3	81	2	3	77	4	3	81	0	3	81	0	3	79	2	3	80	1	3	80	1	3	27	.4
	II	2	59	0	2	54	5	2	53	6	2	57	2	2	57	1	2	55	3	2	58	0	2	58	0	2	58	0	2	28	.9
Granite	I	14	352	6	14	346	12	14	348	8	14	337	18	14	346	15	14	338	23	14	347	15	14	357	6	14	353	11	14	25	.9
	II	11	269	11	11	265	14	11	268	11	11	264	9	11	254	19	11	259	18	11	273	5	11	273	9	11	266	15	11	24	1.1
Nebo	I	4	96	1	4	96	1	4	90	3	4	90	2	4	88	3	4	88	3	4	90	2	4	86	6	1	20	0	3.7	23	.6
	II	4	98	1	4	98	1	4	92	5	4	98	0	4	97	1	4	95	4	4	92	9	3	73	2	1	24	0	3.5	24	.7
Totals	I	21	531	7	21	524	14	21	519	13	21	504	24	21	515	18	21	507	26	21	516	19	21	523	16	18	453	12	21	25	.8
	II	17	426	12	17	417	20	17	413	22	17	419	11	17	408	21	17	409	25	17	423	14	16	404	11	14	348	15	17	25	1.0
	Both	38	957	19	38	941	34	38	932	35	38	923	35	38	923	39	38	916	51	38	939	33	37	927	27	32	601	27	38	25	.9

[a] Average number of students present per class per filmstrip.
[b] Average number of students absent per class per filmstrip.

Table 27

Number of Classrooms, Students Present, and Students Absent for Each Practice Test

District	Experimental Group	PRACTICE TEST 1			2			3			4			5			6			7			$\bar{X}$		
		Classes	Students Present[a]	Students Absent[b]	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent	Classes	Students Present	Students Absent
Cache	I	3	81	2	3	83	0	3	81	1	3	77	4	3	81	0	3	80	1	3	81	0	3	27	.5
	II	2	56	3	2	59	0	2	59	0	2	56	3	2	56	2	2	56	2	2	54	4	2	28	.5
Granite	I	14	350	8	14	343	13	14	348	13	14	347	18	14	343	22	14	346	20	14	347	22	14	25	1
	II	11	263	17	11	264	16	11	253	19	11	264	11	11	260	19	11	264	17	11	263	18	11	24	1.5
Nebo	I	4	93	3	4	85	7	4	90	[illegible]	4	87	5	4	85	7	4	87	5	0	0	0	4	22	1.2
	II	4	97	2	4	94	3	4	96	2	4	98	1	4	99	1	4	94	7	2	50	2	4	24	.5
Totals	I	21	524	13	21	511	20	21	590	16	21	511	27	21	509	29	21	513	26	17	428	22	21	25	1.5
	II	17	416	22	17	417	19	17	408	21	17	418	15	17	415	22	17	414	26	15	367	24	17	24	1.3
	Both	38	940	35	38	928	39	38	998	37	38	929	42	38	924	51	38	927	52	32	795	46	38	25	1.4

[a] Average number of students present per class per filmstrip.
[b] Average number of students absent per class per filmstrip.

the classrooms and brought observational reports back to direct revisions and future filmstrips.

Results from the evaluations of Filmstrips #3 through 9 are shown in Table 28. The first section of the form asked teachers to use a 4-point scale to agree (1) or disagree (4) with positive comments about the filmstrips. The mean response to 11 statements was 1.8 (between agree, 2, and strongly agree, 1). The most positive teacher reaction was received toward Teacher Involvement. Teachers felt their involvement was clearly defined, easy to accommodate, and appropriate. Although no statement received negative feedback (disagree, 3, or strongly disagree, 4), teachers felt less positive ($\overline{X}$ = 2.28) about the filmstrip length than any other item. Some teachers did feel the filmstrips were too long because they took time from other work, but the teachers agreed that students were not bored and enjoyed watching the filmstrips.

The second section of the filmstrip evaluation asked short-answer questions. Results in Table 28 showed that teachers perceived a transfer of student test-taking skills to other subjects, spent minimum preparation time (11.9 minutes per filmstrip), thought that students learned most of the concepts (84%), taught the filmstrips themselves, and used supplemental material 38% of the time.

Additional comments solicited from teachers indicated that the red "highlight" was not an effective method to emphasize words, filmstrips were too long, more student practice was needed, and "elimination " skills were not taught thoroughly.

Practice tests. Feedback on practice tests was collected from teachers by phone and through written comments placed on an

Table 28

SUMMARY OF FILMSTRIP EVALUATIONS*

FILM-STRIP #3	FILM-STRIP #4	FILM-STRIP #5	FILM-STRIP #6	FILM-STRIP #7	FILM-STRIP #8	FILM-STRIP #9	AVERAGE FOR ALL FILM-STRIPS	FILMSTRIP EVALUATION QUESTIONS
Average Rating on Scale of 1=strongly agree to 4=strongly disagree								
								Filmstrip
1.97	2.44	2.31	2.54	2.25	2.06	2.38	2.28	1. The length was appropriate.
1.82	1.88	1.94	1.77	1.97	1.73	1.70	1.83	2. The story line was entertaining to the students.
1.59	1.47	1.81	1.74	1.53	1.73	1.58	1.63	3. The content addressed skills the students need to learn.
1.80	1.91	2.03	1.91	2.00	1.97	1.70	1.90	4. The figures and printing on the filmstrip were clear.
2.00	1.56	1.78	1.91	2.08	1.70	1.58	1.80	5. The dialogue was audible.
2.09	1.70	1.78	1.85	1.79	1.73	1.66	1.80	6. The filmstrip turner was able to move with the narrated page.
								Teacher Involvement
1.82	1.27	1.56	1.57	1.44	1.43	1.45	1.50	7. The teacher was properly cued to stop the tape.
1.73	1.62	1.75	1.71	1.66	1.63	1.64	1.68	8. The amount of Owl/teacher interaction was appropriate.
1.87	1.56	1.66	1.91	1.61	1.52	1.55	1.67	9. The tasks required of the teacher were easy to accomplish and defined clearly.
								Student Materials
1.73	1.97	1.75	2.00	2.05	1.76	1.77	1.86	10. The student practice was sufficient for students to apply the concepts they learned through the filmstrip.
1.84	2.03	1.81	1.94	2.11	1.79	1.87	1.91	11. The practice exercises were of the appropriate difficulty level.
1.84	1.76	1.83	1.89	1.86	1.73	1.72	1.80	TOTAL (AVERAGE FOR FIRST ELEVEN QUESTIONS)
Questions Answered Yes/No or in Minutes								
59%	74%	81%	79%	84%	69%	90%	76%	1. Have students applied test-taking skills to other subjects? (Percentage answering "yes")
25.15 (24.29)	12.73 (9.81)	8.75 (4.21)	9.41 (5.61)	9.55 (6.60)	8.38 (4.75)	9.27 (6.17)	11.89 (8.78)	2. How long did it take to prepare to teach this filmstrip? [Average and (standard deviation) in minutes]
09%	24%	13%	10%	33%	15%	11%	16%	3. Were there any concepts presented in the filmstrip that were not learned by your students? (Percentage answering "yes")
94%	94%	100%	100%	94%	96%	93%	96%	4. Were you the teacher for the filmstrip? (Percentage answering "yes")
42%	32%	37%	48%	42%	28%	39%	38%	5. Did you use the pictures that accompany the filmstrip? (Percentage answering "yes")

OPEN-ENDED COMMENTS IN RESPONSE TO
"#6: If you have any additional comments, please write them on the back of this form."

(Only comments made by 5% or more of the teachers for a given filmstrip are recorded.)

Filmstrip #3

None

Filmstrip #4

Red highlighting doesn't show up well.	29%
Too long.	15%

Filmstrip #5

Red highlighting doesn't show up well.	16%
Too long.	16%

Filmstrip #6

No #4.	23%
Too long.	25%
Teacher needs helper.	8%
More examples needed for identifying sounds.	8%

Filmstrip #7

More practice needed.	13%
Too long.	16%
Red highlighting doesn't show up well.	8%
Concept of "eliminate" difficult for children to learn.	22%
Boring for children.	8%

Filmstrip #8

None

Filmstrip 9

Too long.	13%
Error at 2nd stop: "Bob" should be "Tom."	13%
Red highlighting doesn't show up well.	10%

*Note: Filmstrips #1 and #2 were not evaluated using this form.

identification sheet accompanying returned test forms. Usually, feedback was specific to the district's format or the content of the three levels. In general, teachers felt that the content of the lowest level of the three test formats (SAT, CTBS, and ITBS) and the format of the ITBS was too difficult. Since the intent of administering practice tests was to give the students exposure to all facets of the reading subtest, modifications in the practice test format were not made. However, in response to the feedback, the content of the lowest levels was made easier so that low-achieving students could experience more success before taking the district test, which would probably be very difficult for them. Teachers also noted that later tests were too long (20-30 minutes). Because one objective was to prepare students to take typical standardized tests (which are often 30 minutes long per subtest), the length was not adjusted.

Reinforcement procedures. After the first reinforcement session, informal comments from some teachers suggested that procedures were difficult for the teachers to explain and for the students to understand. Project staff visited those teachers (see Table 20) to model for teachers while reteaching the process to students. During subsequent sessions, teachers reported that students sometimes became upset when they did not earn points. Teachers were told to encourage students to work harder on the next test. Since the number of points awarded to students was a function of the previous test score, students rarely missed getting points on consecutive tests. On the occasion that points were not earned on successive tests, teachers were told to lower the "TO BEAT" score enough for the student to earn a point. All modifications to the original plan were made to increase reinforcing effects of the points and in no way jeopardized the research design nor outcome data.

Project evaluation. After the final student data were collected, teachers in Experimental I and II groups were mailed a Project Evaluation Form (see Appendix F). Teachers responded to 39 statements using a 5-point Likert scale to indicate agreement (1) to disagreement (5). Statements concerned filmstrips, practice tests, contact and communication with project staff, data collection procedures, general impressions, reinforcement procedures (Experimental Group I only), and the spring teacher training workshop (Experimental Group I only). The results of the project evaluation are presented in Table 29 by experimental group. Teachers in both groups had similar attitudes with a mean agreement score of 2.1. Teachers felt more positive toward filmstrips (1.6) than the practice tests (1.9) or the reinforcement procedures (2.9).

Before returning the completed form to USU, each teacher was contacted by phone by project staff. Teachers were asked to add verbal comments to explain their responses to statements in the five areas (seven for Experimental Group I) listed above. These comments are presented in Table 30. In general, verbal comments indicated positive attitudes toward the filmstrips, practice tests, and the project as a whole, and negative attitudes toward the reinforcement procedures and the filmstrip length. The teachers made several suggestions for project improvement. The most frequent suggestions were to provide more student practice on filmstrip concepts, increase the percentage of total filmstrip time spent on reading comprehension, and include skills for math tests in the instructional sequence.

Support and quality of teachers. The degree of project implementation very likely depended to some degree on the support that teachers showed for the project and the quality with which they

Table 29

RESULTS FROM TEACHER EVALUATION: PROJECT COMPONENTS

MEAN ATTITUDE SCORE AND STANDARD DEVIATION

PERCENT OF TOTAL RESPONDENTS

E1		E2		Total			Strongly Agree		Neutral		Strongly Disagree	No Data
Mean	S.D.	Mean	S.D.	Mean	S.D.	Filmstrips	1	2	3	4	5	
1.71	.56	1.81	.64	1.76	.59	1. Instructions for teachers were complete and easy to follow	31.6	60.5	7.9	0.0	0.0	
1.66	.91	1.70	.69	1.68	.81	2. The filmstrips were easy to implement in the classroom	47.4	42.1	5.3	5.3	0.0	
1.28	.56	1.23	.56	1.26	.55	3. The concepts taught in the filmstrips were important for students to learn	78.9	15.8	5.3	0.0	0.0	
1.85	.91	1.75	.58	1.81	.78	4. The filmstrips taught the concepts adequately	34.2	52.6	5.3	5.3	0.0	2.6
1.76	.53	1.64	.70	1.71	.61	5. The students enjoyed the filmstrips	36.8	55.3	7.9	0.0	0.0	
1.71	.88	1.67	.98	1.69	.88	6. I plan to use the filmstrips in future classes	50.0	28.9	10.5	5.3	0.0	5.3
1.76	.94	1.64	.93	1.71	.93	7. The filmstrips were worth the time and effort required	52.6	31.6	7.9	7.9	0.0	
1.68	.48	1.60	.5	1.63	.49	Total Filmstrip Component: Items 1-7						
						Practice Tests						
1.80	.68	1.58	.62	1.71	.65	8. Directions to students were complete and easy to follow	39.5	50.0	10.5	0.0	0.0	
2.00	.10	1.76	.83	1.89	.98	9. Tests were easy to implement in the classroom	39.5	42.1	10.5	5.3	2.6	
2.47	1.20	2.17	.81	2.34	1.04	10. The test items were appropriate in content and and difficulty	15.8	55.3	13.2	10.5	5.3	
2.05	.82	2.00	.79	2.02	.80	11. The tests adequately prepared the students for standardized testing	21.1	57.9	15.8	0.0	2.6	
2.04	1.07	1.81	1.04	1.94	1.05	12. I plan to use the practice tests in the future	39.5	36.8	10.5	7.9	2.6	
2.61	1.07	2.43	1.09	2.54	1.07	13. Students enjoyed taking the practice tests	13.2	42.1	23.7	13.2	5.3	
2.19	1.07	1.82	.88	2.02	.99	14. The practice tests were worth the time and effort required	50.0	34.2	15.8	0.0	0.0	
2.20	.79	2.00	1.00	1.90	.90	Total Practice Test Component: Items 8-14						

Table 29 (cont'd)

Results from Teacher Evaluation: Project Components

MEAN AND STANDARD DEVIATION							PERCENT OF TOTAL RESPONDENTS					
						Contact and Communication	Strongly Agree		Neutral		Strongly Disagree	No Data
E1		E2		Total								
Mean	S.D.	Mean	S.D.	Mean	S.D.		1	2	3	4	5	
1.88	.68	1.47	.78	1.65	.74	15. The USU contact person kept me well informed	50.0	34.2	15.8	0.0	0.0	
7.57	.59	2.18	.75	1.83	.73	16. I was able to reach my USU contact person and felt comfortable in doing so.	34.2	44.7	18.4	0.0	0.0	2.6
1.57	.81	2.05	.83	1.78	.84	17. My needs were responded to in a reasonable amount of time	44.7	34.2	18.4	2.6	0.0	
1.14	.39	1.52	.62	1.31	.52	18. The contact person listened and responded to my feedback	71.1	26.3	2.6	0.0	0.0	
1.40	.51	1.90	.67	1.60	.62	Total Contact and Communication Component: Items 15-18						
						Data Collection						
2.04	.97	1.64	1.00	1.86	.99	19. The observation during testing was non-disruptive	39.5	47.4	2.6	7.9	2.6	
2.23	.89	1.82	1.07	2.05	.98	20. I would not mind having observers again in similar project	31.6	42.1	18.4	5.3	2.6	
2.42	1.00	2.05	.92	2.26	1.08	21. Students enjoyed responding to the student attitude measures on Friday	23.7	42.1	18.4	15.8	0.0	
2.2	.80	1.80	.98	2.00	.81	Total Data Collection Component: Items 19-21						
						General Impressions						
1.95	.81	2.17	.86	2.05	.84	22. The requirements for participation in the study were clearly outlined	26.3	47.4	21.1	5.3	0.0	
2.14	.91	1.94	.83	2.05	.87	23. The benefits were worth the investment of time	26.3	50.0	15.8	7.9	0.0	
2.33	.80	1.82	.88	2.10	.86	24. The project was enjoyable for students	23.7	50.0	18.4	7.9	0.0	
1.95	.92	1.82	.72	1.89	.83	25. The project benefited students' test-taking ability	34.2	47.4	13.2	5.3	0.0	
2.47	.93	2.11	.70	2.21	.84	26. The project enhanced students' performance in other areas	13.2	52.6	23.7	10.5	0.0	
2.14	.91	2.05	.66	2.10	.80	27. The project was realistic in scope	15.8	65.8	13.2	2.6	2.6	
1.95	.62	1.58	1.16	1.78	.96	28. I am glad that I participated	47.7	42.1	5.3	5.3	2.6	
2.23	1.13	2.10	.79	2.13	.99	29. The fall workshop adequately prepared me for the tasks expected	28.9	39.5	23.7	5.3	2.6	
2.09	1.00	1.60	.63	1.88	.89	30. Taking tests was less anxiety-provoking for students because of the project	34.2	44.7	7.9	7.9	0.0	5.3
2.10	.78	1.90	.48	2.00	.67	Total General Impressions Component: Items 22-30						

Table 29 (cont'd)

Results from Teacher Evaluation: Project Components

MEAN AND STANDARD DEVIATION			PERCENT OF TOTAL RESPONDENTS					
		Reinforcement	Strongly Agree		Neutral		Strongly Disagree	No Data
E1								
Mean	S.D.		1	2	3	4	5	
3.14	1.63	31. The reinforcement procedures were easy for students to understand	14.3	.19	14.3	43.8	9.5	
2.71	1.19	32. The reinforcement procedures were easy for the teacher to implement	14.3	33.3	28.5	14.3	9.5	
2.85	1.09	33. Students worked hard to earn more than their "to beat" score on the test	9.5	28.6	28.6	23.8	4.7	4.7
2.71	1.19	34. Students enjoyed the reinforcement procedures	14.3	38.0	14.3	28.6	4.7	
2.90	1.41	35. I plan to use the procedures for reinforcement in the future	24.0	14.3	24.0	24.0	14.3	
2.86	.18	Total Reinforcement Component: Items 31-35						
		Spring Workshop						
2.8	1.00	36. Workshop materials were clear and helpful	48.0	24.0	19.0	0.0	9.5	
2.24	1.41	37. Workshop was appropriate in length	38.0	33.3	9.5	4.8	14.3	
2.28	1.42	38. Information gained from the workshop(s) was worth the amount of time required	38.0	28.6	14.3	4.8	14.3	
2.15	1.46	39. As a result of the workshop, I was a better test administrator	48.0	19.0	4.8	14.3	9.5	4.8
2.10	1.30	Total Spring Workshop Component: Items 36-39						

BEST COPY AVAILABLE

Table 30

Verbal Comments from Teachers on Project

Teacher ID	Filmstrips	Teacher ID	Practice Tests
01	All too long.	01	Good except my kids would have benefited from more practice - less film.
02	Too long.	02	Pretty good.
03	Liked them basically; red color bad; too long but very good at catching kids - they loved them, will use again.	03	At first too easy, then too hard; worthwhile; got kids used to new format.
04	Kids enjoyed the films; a little bit too long.	04	Too long; kids were tired of tests; tests were too hard but after taking the ITBS - I understand why it was so hard - but maybe it would be better to make it hard for real test and easy all year - during practice; just some minor problems with items.
05	I liked them for most part; kids enjoyed; sometimes I couldn't understand characters; especially helped to teach elimination.	05	They got too hard too fast; I understand why it was hard - especially when I saw the real ITBS; tests came too fast at end and kids got tired of them; too long at end.
06	Really enjoyed program; we had a long break in beginning after we had explained the program; should have been more consistent in time line; our materials would sit in post office for 3 days; the end was awful - too crammed together; I will use materials from beginning - spaced throughout year.	06	Sometimes directions were typed wrong; once I went through one set - I knew how to give all; students enjoyed them until the end; very pleased with program; ITBS practice test didn't do anything - to prepare kids - elementary.
07	Part on guessing - deduction was important but was presented too quickly, students need more practice; too close together - students seemed tired of program near the end; did not show #9.	07	Too close - space but over year - should be viewed as part of curriculum; plan to use all practice tests next year - she will have some students in third grade.
08	Sometimes too long, but will use again.	08	No problem.
09	Great.	09	Some better than others, a bit of a pain but did prepare kids.
10	Red ink bad; made good points.	10	Were needed.
11	Enjoyed; no additional comments.	11	Covered all the things kids needed.
12	-	12	-
13	-	13	-
14	Content was excellent; red lettering poor; need to divide some in half.	14	Format good, easy to administer, provided good practice.
15	OK but too long (kids only last 15 min.); divide into one topic at a time.	15	Too hard, one child quit. Need to be easy to build confidence.
16	Content excellent; need to present one concept/film (10 min./day).	16	Need greater differentiation, not easy enough for the slow kids.
17	Well done, but red letters bad. OK for fast kids, too long for slow kids.	17	Too many, kids tired by end of project. Spent too much practice on easy concepts and not enough on hard concepts (reading comprehension).
18	Very enjoyable except #7 - kids thought it was boring (try color).	18	Kids did very well and enjoyed them.
19	Well done, some long.	19	Some of the sounds were difficult.
20	-	20	Fine.
21	OK, went well - see evaluation forms.	21	No problem.
22	Pretty good; too long; spaceman great.	22	Fine.
23	-	23	Fine, good for kid.
24	At first it was shaky; once in a while teacher would not know when to stop - but better at the end with beeps; very clever.	24	Kids enjoyed checking tests - they thought they were smart; checking was really a reinforcement.
25	Kids enjoyed the characters; I would change red; it was good that kids could react to characters.	25	Didn't like the mistakes - I felt I had to proof-read each test; kids did like to take tests, they always did well.
26	Easier if teacher could work program by herself - hard to get; students really enjoyed animals throughout; I still used board and red showed up OK.	26	Sometimes too difficult for low group; pictures were hard to discriminate.
27	Kids really liked them; red was problem.	27	Medium and high were right; your tests were way too hard and it frustrated them; most of low students did not read actual ITBS items - they finished very quickly; a little long; kids were a lot more relaxed for ITBS - they knew exactly what to do for reading; during math, kids seemed confused and more nervous.
28	-	28	-
29	-	29	-
30	-	30	-
31	-	31	-
32	-	32	-
33	Excellent; red bad color.	33	Very adequate.
34	Better once time to respond more accurate; kids enjoyed.	34	Enjoyed; weren't too hard, level appropriate.
35	Red bad color; occasional muffled sound; content good, kids understood.	35	Great. A bit hard for Distar kids.

Table 30 (continued)

Teacher ID	Contact with Project Staff
01	Fine.
02	Once we got in contact, things were faster.
03	Really good.
04	-
05	-
06	Average - few mixups on materials not arriving; would like to have had more contact.
07	We got behind and felt pressured.
08	Great.
09	Fine.
10	The staff did all they could.
11	Good.
12	-
13	-
14	Adequate (team not familiar with the classroom - "ivory tower syndrome").
15	Good.
16	Very good.
17	Good.
18	It has been delightful (good PR).
19	Great, very patient.
20	O.K.
21	Good.
22	Fine.
23	At first hard to catch the staff, but OK.
33	Good.
34	Instructions clear, never any questions.
35	Very adequate.

	Data Collection
01	Great.
02	They started laughing at one of the kids and he knew it--other than that, O.K.
03	Really good.
04	TW test was discouraging after 4 days of testing; I wasn't bothered at all by observers.
05	Didn't bother us.
06	I gave the wrong test; students did not notice observers; students were tired of testing by Friday.
07	Students did not notice the observers; students enjoyed test-wiseness; students liked test administrator.
08	Observers talked and distracted kids.
09	Caused some disruption but no big deal. Would still prefer no observers.
10	Not even aware of them.
11	Fine.
12	-
13	-
14	Fine, they were very quiet and didn't bother children.
15	Test-wiseness and attitude was too rushed on Friday (not enough time).
16	No problem, except one girl sitting close could hear tape (watched observers).
17	With proper introduction, it went well, but by Friday (test-wiseness) kids were too tired.
18	Sat in the front of the room and were somewhat disruptive. Children were very aware they were there.
19	Very quiet.
20	Fine.
21	No problem - the letter panicked everyone but was fine.
22	Good.
23	Fine - were worried but fine.
24	Very good observers; did not notice their presence.
25	I didn't know they were there; observers weren't trained on test administration; kids were a little confused because observer used different language.
26	-
27	Absolutely no disruption - very prepared.
28	-
29	-
30	-
31	-
32	-
33	Nice people.
34	No problem; felt it would be better with professional observers to give test-wiseness test.
35	Totally unobtrusive.

Teacher ID	General Impressions of the Project
01	Definitely worthwhile.
02	Glad she did it.
03	Liked program; really need to teach this; kids had good feeling.
04	Spaced more throughout year; once a month for each activity; took too much time in a short span of time; I will do it again - but space it throughout the year.
05	Project did help with ITBS - they learned format; kids did better on reading than math and spelling; after ITBS kids realized why practice tests were so hard; project took a lot of time at end - that was hard.
06	Really impressed with program; concepts in later filmstrips needed more development and practice and T's didn't have time to supplement - and students really need to know about guessing; W.S. in fall - I didn't really know what I was going to do when I got back to my class - your visit was important.
07	Still confused about what we were to do; wasn't clear on what to bring to workshop #1.
08	Enjoyed, worthwhile, kids more relaxed.
09	Kids did feel more comfortable but maybe too much! Generally worthwhile, would use films again but fewer practice tests.
10	Can't answer until data in. Would rather it to be 1 unit, not so dragged out.
11	They seemed prepared for reading but social studies and science threw them. Kids disappointed it wasn't exactly like Owl.
12	-
13	-
14	Kids were very relaxed this year, easiest administration of SAT in 10 years of teaching, entire project easy to plan, prepare and administer for teacher. Need to use 20-30 minute blocks of time rather than 45-60 minute blocks, because you lose children after 20-30 minutes.
15	Need to work with teachers in planning the study. After "you" people have been out of the classroom for two years, you're "no good". "It is like England trying to rule the colonies." Sometimes toward the end there was a lack of consideration for the students.
16	Although it was hectic at the end, the training really showed during the test. Easiest administration of SAT I've had in 8 years. Really helped our English as second language kids who otherwise would have been wiped out by this experience. "Wished we would have had training in math."
17	Have 2nd grade classroom teachers on planning staff and consult with teachers as you go. Plan the same type of instruction for math.
18	The test-taking skills have generalized and concepts such as learning to eliminate have carried over. Kids seem better prepared to cope.
19	Excellent preparation for the test (kids really learned to proofread and use these skills with other assignments).
20	Later films best; worthwhile.
21	Nothing additional.
22	Need to see results - if they did better it was worthwhile - were some skills they learned for other things.
23	Some films too long; pushing too much.
24	-
25	-
26	A whole outline to show us what to do - title page or table of contents; surprised how relaxed kids were for ITBS reading - kids were apprehensive about math - didn't know how to do items; fall workshop didn't really explain the scope of the program. I knew what was expected in the classroom but didn't know how far to extend into classwork; project was enjoyable - only burden was getting behind.
27	Project was worthwhile - kids learned how to take tests; I will use it again.
28	-
29	-
30	-
31	-
32	-
33	Children and I looked forward to it.
34	Later filmstrips super valuable esp. "Eliminate" concept. Carried through to other areas. Still thinks low students stay low.
35	Great benefits, kids more comfortable.

BEST COPY AVAILABLE

Table 30 (continued)

Teacher ID	Reinforcement (Exp. I Only)
01	My kids not impressed.
02	Don't know for sure how much kids got out of it but excited to take them home. Never got hanging board.
03	I just couldn't get them to understand they were proud of charts but confused. Would use charts but easier to fill in next time.
04	Very negative because it was too time consuming; kids couldn't do it without my help; it seemed to be a chore; better for 3rd and 4th grade.
05	Kids liked it; after you showed them how - they had no trouble; hard for slow students who didn't get many points; more reinforcing for high students.
06	-
07	Students didn't ever really understand procedures even at end; hard for teachers to explain to students; colorful - students liked chart; discouraging for low students - too many "no points" - students should get at least one point each time.
08	Good but not great - too many kids and lots of cheating.
09	Yuk, hating having them circle right answers.
10	The pits - didn't mean anything to kids.
11	A hassle, would rather mark charts up and down, not across. Didn't have space in room - kids liked coloring.
14	Didn't care for it, too bulky and cards fell apart, children never spent any free time with cards, only when instructed to after test.
15	Didn't work, wasn't reinforcement (to beat was too high).
16	Scores not low enough, all excited to take them home.
17	Not very reinforcing, need to attach points to extra recess, etc.
18	Kids were proud of their charts (took them home).
19	Kids liked it.
20	OK.
21	Good except when kids got 0's. Liked coloring.

Teacher ID	Spring Workshop (Exp. I Only)
01	Fine.
02	Good, kids loved going back over questions after testing.
03	Good, excellent.
04	Very informative, excellent; without W.S., I would not have taken test or analyzed the test - I know I should do it anyway but I wouldn't have.
05	Helpful to me; I wasn't so shocked when I gave the test; I could really explain to kids that some items were hard and no one expected students to get them all right.
06	Super, really good; we could have spent 2 days on it.
07	Most beneficial of all; really prepared T for giving test; ideas presented were not in manual; taking the test taught me what to teach kids for taking test.
08	Fine, learned a lot.
09	Good.
10	Left in middle cuz daughter had baby - maybe too minute in detail.
11	Good, learned a lot of new incorporations.
12	-
13	-
14	Unnecessary (could be done in 30 minutes).
15	"Defeated purpose of your study" (confused the question of student improvement by changing teachers - should let teachers do it normally).
16	Was a little long, but learned three valuable lessons: (1) stand or sit in front of room rather than roaming around; (2) makes notes and observations of students during the test; (3) carefully go over directions and EXPLAIN IN DETAIL.
17	Waste of time.
18	Didn't attend.
19	Didn't attend.
20	Interesting.
21	Clear and useful.

BEST COPY AVAILABLE

administered the student training. Staff members who had personal contact with the teachers through the project rated each teacher on a scale of 1-3 in both quality and support before any other project data were collected. Guidelines for the criteria for each of the ratings were as follows:

SUPPORT FOR PROGRAM:

3 - Seldom complained, receptive to necessary change, attended workshops, eager to cooperate, punctual with materials, very positive over the phone and when observed.

2 = Occasionally complained, somewhat resistant to change, partial attendance at workshops, cooperated, generally punctual with materials, occasionally apathetic but not antagonistic when observed.

1 = Always complaining, very resistant to change, failed to attend workshops, little cooperation, general negative attitude over the phone and when observed.

QUALITY OF IMPLEMENTATION:

3 = Always on schedule with filmstrips and tests, returned materials in proper order, no major deviations from implementation (test administration, reinforcement, etc.), followed directions when observed.

2 = Close to schedule with filmstrips and tests, returned almost all materials (some with mistakes), moderate deviations in implementation (changed "TO BEAT" scores, etc.), classroom observations were fair.

1 = Seldom on schedule, missing materials, materials received had major errors (did not use reinforcement charts, etc.), observations poor.

Results, summarized in Table 31, indicate that in general, teachers demonstrated strong support for the project ($\overline{X}$ = 2.53) and implemented the components in a high quality manner ($\overline{X}$ = 2.42).

Table 31

Mean Ratings Given Teachers for Support and Quality

	Percent of Teachers Selected for Each Rating			Mean Rating					
				District			Experimental Group		
	3	2	1	Cache	Granite	Nebo	E1	E2	Total
Support	57.9	36.8	5.3	3.0 n = 5	2.5 n = 25	2.4 n = 8	2.6 n = 21	2.5 n = 17	2.5 n = 38
Quality	57.9	26.3	15.8	2.8 n = 5	2.4 n = 25	2.3 n = 8	2.5 n = 21	2.4 n = 17	2.4 n = 38

INSTRUMENTATION

To determine the effects of the experimental treatment, a variety of dependent variables were considered which provided information about both students' and teachers' performance during the standardized test administration. Data collected included standardized achievement test scores from the test used in each district and a variety of locally developed measures which examined such variables as student and teacher on-task behavior during the testing, the quality of test administration, student and teacher attitude towards testing, and student test-wiseness skills. The remainder of this section provides a brief description of each of these dependent variables.

Standardized Achievement Tests

The major objective of this project was to provide an intervention which would result in more valid test scores. Consequently, it is logical to examine the standardized achievement test scores of children in the experimental and control groups to determine whether there are differences between the scores. If the experimental treatment resulted in more valid scores, then one would expect children in the experimental treatment to score differently on the average than children in the control groups. Because each district included in the study was using a different standardized achievement test in their district testing program, it was necessary to convert the scores to a standard metric before including them in the analysis. Z score transformations (Glass & Stanley, 1970, p. 87) were computed for each student's score using the following formula:

$$\overline{X}_{ij} - \overline{X}_{.j} \div SD_{.j} = Z_{ij}$$

where i equals the ith student and j equals the district (either Granite, Nebo, or Cache). Z scores were computed within each district. In other words, the mean and standard deviation of all of the participating students' scores in Granite were computed and used in conjunction with each individual student's score to compute a Z score. Since each district had approximately the same number of Experimental Group I, Experimental Group II, and control students, this procedure yielded a score which could be combined in one total analysis even though districts used different standardized achievement tests.

Each standardized achievement test was administered by the classroom teachers in the individual districts which is the procedure normally followed in each of the three districts participating in the project. Granite and Nebo districts administered the tests the week of March 29th to April 2nd, and Cache District administered the test the week of April 5th to 9th. All tests were scored by the respective publishing companies and returned to the district offices, who then made scores available to the research staff. Experimental Group II and Control Group teachers were instructed to follow the normal procedures in their district for administering the test.

Students in Cache District completed the most recent version of the Comprehensive Tests of Basic Skills, Form U/Level D, Grades 1.6-2.9 (CTBS, 1981). The battery is made up of 10 subtests in six content areas (see Table 32) of which three focus on reading: word analysis, vocabulary, and reading comprehension. This particular version of the CTBS was piloted in 1979 and standardization was conducted in the fall of 1980 and spring of 1981. Reliability data are available in the test coordinator's handbook.

Students in the Nebo District completed the Iowa Tests of Basic Skills, Form 7/Level 8 (ITBS, 1980). The battery is made up of 15 subtests in seven skill areas (see Table 32) of which five focus on reading: vocabulary, word analysis, picture comprehension, sentence comprehension, and story comprehension. According to Buros Mental Measurement Yearbook (8th edition),

Table 32

Standardized Test Formats

TEST	SUBTESTS	Teacher Directed	Student Directed	# Items	# Minutes
CTBS 1981	Word Attack	ⓧ		40	38
	Vocabulary		X	25	19
	Reading Comprehension		ⓧ	25	28
	Spelling		X	25	17
	Language Mechanics		X	20	15
	Language Expression		X	25	27
	Mathematics Computation		X	20	18
	Mathematics Concepts	X		30	33
	Science	X		25	28
	Social Studies	X		25	28
ITBS	Listening	X		32	16
	Vocabulary		X	20	14
	Word Analysis	ⓧ		57	20
	Reading Comprehension				
	Pictures		ⓧ	23	12
	Sentences		X	16	7
	Stories		X	28	15
	Language Skills				
	Spelling	X		29	13
	Capitalization		X	75	12
	Punctuation		X	68	13
	Usage	X		23	9
	Work Study Skills				
	Visual Materials	X		32	24
	Reference Materials	X		38	25
	Mathematics Skills				
	Mathematics Concepts	X		36	15
	Mathematics Problems	X		24	18
	Mathematics Computation	X		28	22
SAT	Vocabulary	X		37	20
	Reading Part A		X	45	20
	Reading Part B		ⓧ	48	25
	Word Study Skills A	ⓧ		30	10
	Word Study Skills B		X	35	15
	Math Concepts	X		35	20
	Math Computations		X	37	30
	Math Applications	X	X	28	20
	Spelling	X		43	25
	Social Science	X		27	20
	Science	X		27	20
	Listening Comprehension	X		50	35

Circles indicate those subtest scores during which student and teacher on-task observational data were collected.

the split-half reliability coefficients on composite scores and equivalent forms range from .60 to .94. The intercorrelations from the subtests range from .69 to .83 with a median of .76.

Students in the Granite School District used the Stanford Achievement Test, Form A/Primary Level 2 (SAT, 1973). The test is made of 12 subtests in eight content areas with five subtests focusing on reading; vocabulary, reading part A, reading part B, word study skills A, and word study skills B (see Table 32). Standardization of the SAT has been extensive with split-half reliability coefficients generally reported in the high .80s to mid .90s.

Five scores were recorded for each student. Because the student training focused on the content area of reading, three different reading scores were obtained: a teacher-directed test (i.e., a test in which each item was read by the teacher and the timing of the test was paced by the teacher), a student-directed test (i.e., a test where the teacher gives the directions and then gives students a specified time limit to work a number of problems at their own pace without further directions), and the total reading test. Those subtests selected for the teacher-directed and student-directed test in each district are circled in Table 32. In addition to these subtest scores, each student had a total reading, a total math, and a total test score recorded. The rationale for recording student-directed and teacher-directed scores separately was two-fold. First, it was felt by the project staff that the skills taught in the filmstrips were very different for teacher-directed and student-directed tests. Secondly, as will be noted below, the on-task observations of students and teachers took place during the same student-directed and teacher-directed subtests that were included in this analysis. In this way, the relation between on-task behavior and student scores could be observed.

Student and Teacher On-Task Behavior

Student and teacher testing behaviors that are both appropriate and thought to produce more valid scores are frequently outlined in test administration manuals. These behaviors, particularly those of the teacher, are usually specified as "standardized procedures." Adherence to standardized procedures is necessary to achieve comparative, normative data. Unfortunately, few data show that these preferred behaviors actually do influence the validity of test scores. Additionally, even though assumptions are made that teachers follow certain (e.g., standardized) testing behaviors, there is no evidence demonstrating that these behaviors are being displayed. That is, are teachers and students really doing what the teacher's manual and other documents specify as "good practice"? Questions such as "What is 'on-task' during testing?", "What do students and teachers really do during testing?", and "Do certain student and teacher behaviors affect test results in and of themselves?" have not been answered. Two instruments, described below, were developed to gather data about these questions. Specifically, the following questions were addressed:

1. Do teachers follow the directions prescribed in test manuals to establish appropriate environments prior to, during, and after test administration?
2. Do teachers in different experimental groups implement procedures to various degrees depending on treatment conditions?
3. Do students attend to teacher directions and the test items during testing?
4. Do students in different experimental groups attend to tasks in varying degrees depending on the treatment conditions?

Two instruments were devised by project staff to collect data that would describe classroom behaviors during testing. One measure was a checklist of items which were checked off by observers as testing activities were initiated. The other measure was an interval recording system for collecting on-task behavior of students and teachers. Similar versions of both the checklist and the observation recording form had been developed under another project. The next sections describe the original instruments, the preliminary revisions, the pilot tests, and the final revisions.

Original instruments. Initially, the Quality of Test Administration Checklist consisted of a list of activities which were initiated by the teacher and occurred prior to, during, and after test administration. The list was generated from test administration manuals, research on classroom teaching techniques, and textbooks on psychometrics and test administration. Data were collected by pairs of trained observers who checked off items as they were observed during group standardized testing.

The instrumentation to collect on-task data originally included interval recording form and extensive definitions of teacher and student on-task behavior during testing and teacher contact. Data were collected by pairs of trained observers in conjunction with the checklist data. Mean interrater agreement for this version was .88, with a range of .74 to .97.

Pilot test. Both the checklist and the behavioral observation systems and the observer training were piloted for use with this project using a group of 10 graduate students in a research class. The students were trained and then they collected data during testing situations in several second grade classrooms.

Final revisions. As a result of the pilot test, several major changes were made in both instruments. Changes in the checklist included rewording some items to be observable behaviors, adding subjective items that would gauge a general negative or positive climate, and writing exact directions for the observers. Changes in the behavioral observation system involved writing new definitions for "on-task" behavior, distinguishing between teacher-directed and timed tests, and notations for when students finished the timed test. A detailed description of each instrument (as it was used in this study) is provided below.

Checklist. A copy of the final checklist used with this project is included with Appendix G and test statistics are reported in Table 33. The checklist is divided into three sections: teacher behavior before administering the test (16 items), teacher behavior during test administration (15 items), and questions concerning the classroom arrangement and atmosphere (8 items).

In addition to the 31 items which related directly to the quality of test administration as per the teacher's administration manual, other information was collected that was thought to impact on quality of test administration such as disruptive occurrences during the test, noticeable cheating, seating arrangements, and the teacher's use of the aide.

Typically, observers would check some items during their observation and some items after leaving the room. The checklist was always used by pairs of observers during standardized testing. Interrater agreements were computed for each classroom using the equation

$$\frac{\text{Number of Agreements}}{\text{Number of Total Items}} \qquad (1)$$

for an overall mean of .91 (SD = .095).

Behavioral observation. A review of the literature and previous observations contributed to a list of appropriate student behaviors most conducive to producing high levels of attention to academic tasks.

Table 33

Test Statistics on Data Collection Instruments
Developed by Project

	Number of Items	$\overline{X}$	SD	Reliability[a]	Standard Error of Measurement	% Agreement Between Observers $\overline{X}$	% Agreement Between Observers SD
On-Task Behavior	N/A	89.8	11.8	N/A	3.66	90.4	6.0
Quality of Teacher Test Administration	31	51.1	6.3	.82	2.62	90.6	9.5
Teacher Attitude	30	87.7	12.4	.89	3.97	N/A	
Student Attitude	8	12.0	3.6	.80	1.51	N/A	
Test-wiseness	38	21.3	5.0	.75	2.50	N/A	

[a] Reliability estimates computed using Hoyt's measure of internal consistency (Hoyt, 1941; Magnusson, 1967, p. 117).

Definitions of what to consider on- and off-task during testing were derived from this review. Hence, students were considered on-task when looking at the teacher or their test booklet (during teacher-directed or timed subtests) and when following directions. Students were off-task when they displayed any other behavior. A third category of student behavior was observed, "probably on-task", to accommodate those gray areas when observers could not be precise in their "on-task" coding. This situation would occur when the students appeared to be following directions although they were looking away from teacher or test booklet.

Standardized testing procedures listed in the testing manuals and preliminary observations during another project formed the basis for defining teacher on-task behavior. Actions consistent with attending to the students' behavior at all times (while directing the test administration under standardized conditions) are defined as on-task behaviors. For example, during a timed test a teacher is on-task when orally reading directions but is off-task when just talking to the entire class. Essentially, teachers were on-task when reading aloud from the manual or watching the students from the front of the room. Behavior definitions for both students and teachers are summarized in Figure 4.

An interval recording form was used to collect the on-task data on both students and teachers (see Appendix G for a copy of the form and Table 33 for test statistics). Observers were paired for each observation to collect data on five students and one teacher during each observation. Student names were not used on the observation form. Instead, observers randomly selected five students in each classroom and noted a physical characteristic and the type of each student column so that they could move from child to child as quickly as the intervals indicated.

Data recording began when the test administrator started reading the directions and ended when the subtest was completed. Five-second intervals consisted of 3 seconds to observe and 2 seconds to record. Observers watched each child for 4 consecutive intervals or 20 seconds (4 intervals X 5 seconds = 20 seconds) before moving to the next student. Data were recorded on five students and one teacher (six subjects) for a total of 2 minutes (6 X 20 seconds) before repeating the cycle.

Data were placed on the recording form at a signal from a tape recording that indicated when to observe (3 seconds) and when to record (2 seconds). Portable tape players were equipped with earphones for two people to use simultaneously, facilitating interrater agreement calculation. Recording started when the teacher began the directions and observers marked each cell for on-task (1), off-task (0) behavior, or probably on task (-). Each of the five students and the one teacher was observed for 20 consecutive seconds, or four cells, during each 2-minute

Figure 4. Basic Definitions for On-Task Behavior

	STUDENTS		
	1 Definitely On-Task	- Probably On-Task	0 Definitely Off Task
Teacher Directed	Following directions given by teacher with eyes focused on teacher or test booklet	Could be following directions but eyes not focused on teacher or test booklet while teacher reads directions or after students finish item	Not following directions given by teacher or misbehaving out of seat
Timed Test directions	(as above)	(as above)	(as above)
during timing	After test starts until teacher says stop, students must be looking at test booklet or teacher	X	Not looking at teacher or test booklet or misbehaving Out of seat Talking aloud

TEACHER

1 ON-TASK	0 OFF-TASK
During directions or teacher directed items, must be in front of the room. When not reading directions or items, teacher is either looking at students or assisting a student.	Not in front of the room while reading to students from manual. Looking at something other than students during timed test.

block of time. At the end of 2 minutes or once across one row of interval cells, a 5-minute pause gave observers a chance to make notes and locate their position with the first child again before starting the next 2-minute observation. Observers computed percent on-task by dividing the number of "1" marks by the total number of intervals. All computations were checked by a second person and errors adjusted.

Procedures for data collection. Personnel hired to collect data were members of Title I Parent Advisory Councils in Cache, Granite, and Nebo school districts. A total of 22 observers were hired at $5.00 per hour including training, data collection, and travel to schools.

Data collectors were trained simultaneously to administer both the behavioral observation form and the checklist. Training consisted of three segments: (a) practice with videotaped scenes of classroom testing, (b) practice in the classroom during actual testing, and (c) retraining. An outline of the initial training segment conducted from 9:00 to 3:00 on March 26, 1983, is included in Appendix G. The data collectors were kept naive of the experimental design and research questions. Basically, the training sessions led them through each component of the observation and checklist procedure. They rehearsed data collection, used videotaped scenes, and practiced setting up equipment. A list of observation procedures is located in Appendix G.

The schedule of classroom practice and actual data collection is located in Appendix G. Two subtests were observed in each teacher's classroom: a teacher-directed test and a student-directed (timed) test (Table 32 indicates the subtests observed in each district). In all cases, the teacher-directed test was given before the timed test. Observers were assigned by pairs to classrooms. One or two data collectors were designated as substitutes each day. Dates and times for actual classroom observation were randomly assigned across experimental groups and districts (see Table 34 for this breakdown).

Observers were assigned in pairs to classrooms to practice data collection on the first testing day in each district (March 29 in Granite and Nebo and April 5 in Cache). Classroms selected for the practice sessions are listed in Appendix G, and their data were not included in the analysis. Practice was provided on two tests: a teacher-directed and a student-directed (see Table 32 for the subtests in these categories).

Prior to the practice, observers watched one subtest being administered to get a "feel" for the classroom situation and they recorded no data. Observers then recorded behavior on the next two subtests as described above. Data collected during the classroom practice obtained an overall interrater agreement of 86.8 for on-task behavior observations (84.8 with the teacher-directed test and 88.9 with the student-directed test) and 93.9 for the Quality of Test Administration Checklist (see Tables 35 and 36 for a breakdown by district).

A retraining session was held during the afternoon of the practice data collection on Monday. At this time, definitions were clarified, disagreements among observers were solved, and forms were checked by staff personnel for completeness and accuracy.

Actual observations began on Tuesday (March 30 for Granite and Nebo and April 6 for Cache), the second test day and continued through Thursday of the same week. Observers were randomly assigned by different pairs each day to observe both a teacher-directed and student-directed test. These tests were administered consecutively and their order was randomly assigned across teachers. Each day one observer was not assigned to a classroom and was available as a substitute in case an assigned observer did not show up.

Table 34

BREAKDOWN FOR OBSERVATIONS BY NUMBER OF CLASSES

Test/ District	Group	Number of Classes	PRACTICE	ACTUAL OBSERVATIONS						ATTITUDE & TEST-WISENESS			
			Monday	Tuesday		Wednesday		Thursday		Friday			
			8-11	8-10	10-12	8-10	10-12	8-10	10-12	9-10	10-11	11-12	12-2
SAT GRANITE	E1	14	2	2	2	2	2	2	2	4	3	3	4
	E2	11		2	2	2	2	2	1	3	3	2	3
	C	8		2	1	2		1	2	2	3	1	2
	Other		2										
CTBS CACHE	E1	3				2	1			1	1		1
	E2	2						1	1	1	1		
	C	5		1	1		1	1	1	1	1	2	1
	Other		3										
ITBS NEBO	E1	4		1	1			1	1	1	1	1	1
	E2	4				1	1	1	1	1	1	1	1
	C	7	2	2	1	1	1			2	2	1	2
TOTALS		58	9	10	8	10	8	9	9	16	16	11	15

Table 35
PRACTICE DATA COLLECTION
Percent of Interrater Agreement for Quality of Test Administration

District	N	Percent Agreement	SD
Cache	3	93.0	7.0
Granite	6	94.5	6.0
Nebo	2	93.5	4.9
OVERALL	11	93.9	5.6

Table 36
PRACTICE DATA COLLECTION
Percent of Interrater Agreement for On-Task Behavior During Teacher and Student Directed Tests

District	Teacher Directed	Student Directed	Overall
Cache n = 3	82.0 *14.7*	91.0 *9.0*	86.7 *12.1*
Nebo n = 2	75.3 *8.1*	84.5 *2.1*	86.8 *8.4*
Granite n = 6	89.3 *4.4*	89.2 *7.0*	89.2 *5.6*
All Districts	84.8 *9.6*	88.9 *6.9*	86.8 *8.4*

Note. Italicized numbers are the standard deviations.

Substitutes not needing to fill a vacancy computed mean on-task percentages and interrater agreement. Interrater agreements were computed using Equation 1 at 90.6 for observations and 90.44 for the checklist. These data are reported in Tables 37 and 38 by district and summarized in Table 33.

Dependent measures. For the Quality of Test Administration Checklist, the number used in analysis of data is the percent of the 31 items scored as "occurring" in the classroom. For the behavioral observations, percent of time was used in the final statistical analysis. Two percentages were computed separately for both teacher and student on-task behavior: on-task behavior during teacher-directed and during student-directed (timed) tests. Computations of on-task behavior combined percentages of definitely on-task and probably on-task since interrater agreements were comparable whether these scores were separated or combined.

Locally Developed Instruments

As noted above, data were also collected about teacher and student attitude and student test-wiseness. Because no appropriate instruments could be identified in these areas, project staff developed instruments, pilot tested them, and revised them as necessary for use in the project. Table 33 contains some descriptive information including number of items on the three measures, mean and standard deviation, reliability, and standard error of measurement for each instrument. This information will be helpful in interpreting the results of these tests in the Results section. A brief summary of the content and development procedures for each instrument is described below.

Teacher attitude towards standardized tests. It is not uncommon for classroom teachers to feel fairly negative about standardized achievement tests. Although standardized achievement tests can cause many problems, it was our conviction that properly administered and interpreted, standardized

BEST COPY AVAILABLE

Table 37

Actual Data Collection
Percent of Interrater Agreement for Quality
of Test Administration

District	N	Percent Agreement	SD
Cache	10	97.70	3.56
Granite	31	88.96	11.16
Nebo	13	88.38	10.86
OVERALL	54	90.44	9.51

Table 38

ACTUAL DATA COLLECTION

Percent of Interrater Agreement for On-Task Behavior
During Teacher and Student Directed Tests

	District	TUESDAY		WEDNESDAY		FRIDAY		DISTRICT MEAN
		8:00	10:00	8:00	10:00	8:00	10:00	
TEACHER DIRECTED	Cache n = 10	94.5 *1*	95.0 *1*	88.5 *8*	88.0 *3*	90.5 *2*	90.0 *1*	90.8 *4*
	Granite n = 30	84.5 *7*	86.6 *7*	91.7 *4*	92.3 *4*	89.6 *6*	88.4 *7*	88.6 *6*
	Nebo n = 13	85.9 *5*	86.5 *7*	86.1 *4*	82.3 *6*	87.5 *9*	87.9 *8*	86.0 *5*
	ALL n = 53	86.7 *7*	87.7 *7*	89.9 *5*	88.2 *6*	89.3 *5*	88.4 *6*	88.7 *13*
STUDENT DIRECTED	Cache n = 10	92.5 *1*	99.0 *1*	91.5 *4*	95.5 *4*	95.5 *1*	96.0 *1*	94.5 *3*
	Granite n = 31	91.2 *3*	95.2 *4*	96.3 *2*	95.3 *5*	90.8 *7*	94.2 *4*	93.8 *4*
	Nebo n = 13	90.5 *3*	94.3 *4*	92.2 *4*	94.8 *2*	84.5 *11*	78.2 *27*	89.2 *10*
	ALL n = 54	91.2 *3*	95.4 *4*	94.5 *3*	95.1 *3*	90.7 *6*	91.1 *13*	92.8 *6*
	OVERALL n = 107	88.9 *6*	91.5 *6*	92.2 *5*	91.9 *6*	89.8 *6*	89.4 *10*	90.6 *6*

Note. Italicized numbers are the standard deviations.

achievement tests can provide a valuable tool for the educational process. Furthermore, we hypothesized that teachers' attitude toward standardized testing would change as they understood more the purpose of standardized achievement testing, felt that students had been adequately prepared to take the test, and became more skilled in administering the test. An extensive search for a teacher attitude towards standardized achievement tests yielded only one instrument that was reasonably close to what we needed in the project (Beck & Stetz, 1979; Stetz & Beck, 1979). This instrument was used as a basis to refine and develop an instrument in which teachers were asked to respond to Likert-type items in five categories: general opinion, attitude toward administering tests, usefulness of tests, students' feelings about tests, and whether tests should be used more frequently. The total scale consisted of 35 items. The prototype of the instrument was critiqued by project staff members and other testing experts at Utah State University and then pilot tested on an individual basis with four second grade Logan District teachers. Following this pilot test, revisions were made; some items were added and directions were classified; and the test was administered to two classes of 35 teachers who were attending an in-service training program sponsored by Utah State University. Each of these teachers were currently teaching in Granite School District, although none were participating in the test-taking skills project. Item analyses were computed for each of the classes, and point biserials and difficulty levels were used to further refine and improve the test. As noted in Table 33, the final reliability coefficient estimate was .89 and scores were reasonably well distributed. The actual instrument used to collect data on teacher attitude is included in Appendix G along with item statistics from the three groups participating in the study (the format of the questionnaire as it appears in Appendix G has been changed slightly to accommodate the display of item statistics).

This measure of teacher attitude towards testing was delivered to teachers by trained observers who also administered the student attitude form described below. The test was administered on Friday of the week in which the district did standardized achievement testing and then picked up by the observer (see Table 34 for schedule). Each teacher filled out the questionnaire independently (requiring approximately 10 to 15 minutes) during the time the student test-wiseness and attitude measures were being administered. No problems were noted by the observers in collecting the data, and all teachers completed the questionnaire.

Student attitude towards standardized tests. A second major objective of the project was to reduce the anxiety that many students feel during standardized achievement testing and make standardized achievement testing a less threatening and more comfortable experience. No measures for assessing second grade students' attitude towards standardized achievement testing could be located. Therefore, the project developed a measure which was administered by the same people who collected the on-task data during the testing period. These data were collected on Friday of the week in which the standardized achievement testing was done so that the testing experience was still fresh in students' minds (see schedule in Table 34). The actual instrument used is included in Appendix G. The instrument consisted of nine three-point semantic differential type items regarding standardized achievement testing. Directions for administering the test are also included in Appendix G. The person administering the test talked students through each item using a direct instruction mode (defining objectives, giving examples, leading the students through examples, testing them to make sure that they understand, and then proceeding to the test). None of the people administering the test knew which classes were in which experimental group. Appendix G also contains item statistics for each of the items in the test. The reliability estimate of .80 for an eight-item test is quite high and scores were distributed fairly well as shown in Appendix G).

Test-wiseness. Millman, Bishop, and Ebel (1965) defined test-wiseness as "a subject's capacity to utilize the characteristics and formats of the test and/or the test taking situation to receive a high score." According to Millman et al., test-wiseness is "logically independent of the examinee's knowledge of the subject matter." As a part of this project, we differentiated between test-wiseness (strategies that allow a student to get the correct answer on a test even when they have no knowledge of the content being tested) and test-taking skills (mastery of skills that allows a student to demonstrate knowledge that they do have about the content area instead of being confused by strange format or anxiety-provoking experiences). This instrument combined both test-wiseness and test-taking skills.

The instrument was divided in three sections. The first part of the test focused on test-wiseness skills following the outline proposed by Millman, Bishop, and Ebel. Items in the following areas were generated:

1. Eliminate options which are known to be incorrect and choose from among remaining options (deductive reasoning).
2. Choose neither or both of two options which imply the correctness of each other.
3. Restrict choice to options which encompass all of two or more given statements known to be correct.
4. Utilize relevant content information in other test items and options.
5. Select option which is in logical position among an ordered set of options.

6. Consider relevance of specific detail when answering a specific item.
7. Recognize and use specific determiners (e.g., often, seldom, always, never).
8. Recognize and make use of resemblances between the options and an aspect of the stem.
9. When no other information is available, choose the longest alternative.
10. Select option which agrees grammatically with the stem.

The second area was related more directly to elimination and guessing strategies which are a part of the test-taking skills taught by the project. Elimination and guessing are more test-taking rather than test-wiseness skills because they only help the student who has some knowledge about the content being tested. The final section of the test focused on the student's ability to follow directions which are different from what he or she is used to getting. This skill was an important part of what the filmstrips attempted to teacher students.

Two forms of the "test-wiseness" test were developed. Each of these forms was administered to five different individual students and notes were made about where students were having difficulty understanding the test or where the items were not functioning as desired. The two forms of the test were then administered to two classes for each form (four classes in total) and results were submitted to an item analysis program which provided difficulty level, point biserials, and subtest correlations. In addition, student's reading ability was correlated with scores on the test. Using this information, a final version of the test was developed using items from both versions of the pilot test. The final version used approximately half of the items that had originally been developed. Numerous changes in wording, distractors, and arrangement of the items was made during this pilot testing. The final version of the test consisted of 38 items with a reliability estimate of .75 noted in Table 33. The somewhat low reliability estimate is in a part a function of the difficulty level on parts of the test. For example, Part C had an average difficulty level of .863. As Hopkins and Stanley (1981) point out, reliability estimates computed via measures of internal consistency are very sensitive to extreme high or low difficulty levels and are always lower the farther the difficulty level is from .50. A copy of the measure of test-wiseness with the directions for administering the test and selected item statistics for the three groups participating in the project are included in Appendix G.

Accuracy checks on coded data. All data collected with the instruments described above were subjected to accuracy checks. First, *all* computations (including percentages and score sums) were computed twice. Second, data were transferred from one form to another or entered from standardized testing reports to make sure the correct number had been entered in the correct column for the right person. After all data were entered in the master file, frequencies and descriptive statistics (means, standard deviations, minimums and maximums) were computed and checked against possible values.

Summary

Various sources of data were used to examine the effect of the experimental treatment on students and teachers. Standardized achievement test scores, teacher and student attitude towards testing, teacher and student on-task behavior, students' test-wiseness, and the teacher's quality of test administration were all considered. Data from these instruments were combined to form 13 dependent measures which were used in the statistical analyses. The intercorrelations between the 13 dependent measures are presented in Table 39.

The best means of determining what was really being measured by these different instruments, particularly those that were developed locally, is to carefully examine the copies of the instruments included in Appendix G. All but the standardized achievement tests are included in essentially the same format, including directions, in which they were administered during the project. Some spacing changes have been made to accommodate the item statistics reported in Appendix G, but wording and order of items is identical. The reader is encouraged to consider these instruments carefully in interpreting the results reported in Chapter IV.

Table 39

Intercorrelations of Dependent Measures

VARIABLES	Teacher Attitude Towards Testing	Teacher On-Task Teacher Directed	Teacher On-Task Student-Directed	Student On-Task Teacher-Directed	Student On-Task Student-Directed	Achievement Test Reading: Student-Directed	Achievement Test Reading: Teacher-Directed	Achievement Test Math Total	Achievement Test Reading Total	Achievement Test Total Battery	Quality of Test Administration	Student Test-Wiseness	Student Attitude Towards Testing
Teacher Attitude Towards Testing		-.12	-.06	.12	-.003	.01	-.002	-.01	.01	-.004	-.12	-.02	.03
Teacher On-Task Teacher Directed			.48	-.009	.009	-.04	-.05	-.09	-.06	-.07	.35	-.02	-.02
Teacher On-Task Student Directed				.02	-.04	-.02	-.006	.06	-.02	.03	.30	-.03	-.02
Student On-Task Teacher Directed					.25	-.002	.01	.001	.007	.007	.22	.04	-.02
Student On-Task Student Directed						.04	.06	.07	.009	.03	.10	-.01	.05
Achievement Test Reading: Student Directed							.76	.61	.94	.87	.07	.13	-.04
Achievement Test Reading: Teacher Directed								.61	.87	.83	.08	.12	-.04
Achievement Test Math Total									.67	.85	.05	.13	-.03
Achievement Test Reading Total										.93	.05	.13	-.04
Achievement Test Total Battery											.09	.15	-.04
Quality of Test Administration												-.01	-.08
Student Test-Wiseness													-.07
Student Attitude Towards Testing													

CHAPTER IV

RESULTS AND CONCLUSIONS

As described earlier, the research objectives of this project were to:

1. Determine the effectiveness of training materials developed to teach elementary school students test-taking skills, motivate students to do their best on standardized achievement tests, and train teachers in standardized test administration skills.
2. Determine the relationship between scores on standardized achievement tests and students' test-taking skills, teachers' test administration skills, and students' level of motivation.

To provide information about these two objectives, data were collected from 58 classes (containing over 1,400 second grade students) were randomly assigned to one of three groups. Experimental Group I (E1) receiving training in test-taking skills (filmstrips and practice tests), reinforcement procedures, and training in standardized test administration. Experimental Group II (E2) received only the training in test-taking skills (filmstrips and practice tests), and Control Groups (C) received no special curriculum or training procedures related to administering or taking standardized achievement tests. Data were collected from each group about:

1. Teachers' perceptions of the value of the training materials and procedures.
2. Teacher and student attitude towards standardized achievement testing and behavior during standardized achievement testing.
3. Students' scores on the standardized achievement test.

In addition, substantial demographic and implementation data were collected to assist in interpreting the results of the dependent measures described above. A listing of the data collected during the project is contained in Appendix H along with a description of the data file (i.e., variable names, labels, and columns in which data are located). The complete data file is available from the authors. The remainder of this section reports the results of the analyses used to answer the two major research objectives outlined above and uses those results to draw conclusions about the project.

Effectiveness of Training Materials

The degree to which the training materials and procedures were effective in teaching students test-taking skills, motivating students to do their best on standardized achievement tests, and teaching teachers skills in standardized test administration can be judged in terms of teachers' perceptions of the project and the objective data gathered by standardized achievement tests, locally developed instruments, and observers who were uninformed about the nature of the project. The results of the data collection in each of those areas are summarized below.

Teachers' Perceptions

As noted previously in the Implementation Section, most components of the project were viewed very positively by teachers. The filmstrips and practice tests were particularly well received. For example:

- 84.2% of the teachers felt the filmstrips were worth the time and effort required.
- 78.9% of the teachers plan to use the filmstrips next year.
- 94.7% of the teachers felt the filmstrips taught concepts which were important for students to learn.

- 79.0% of the teachers felt the practice tests adequately prepared the students for standardized testing.
- 76.3% of the teachers plan to use the practice tests in the future.
- 84.2% of the teachers felt the practice tests were worth the time and effort required.
- 76.3% of the teachers felt the benefits of the total project were worth the investment of time.
- 73.7% of the teachers felt the project was enjoyable for students.
- 81.6% of the teachers felt the project benefited students' test-taking skills.
- 78.9% of the teachers felt taking tests was a less anxiety-provoking experience for students as a result of the project.

Teachers' perceptions of the procedures for teaching standardized test administration skills were also positive. Seventy-one percent of the participating teachers felt that they were better test administrators as a result of the workshops. Typical comments from teachers concerning the training in standardized test administration were as follows:

- "Very informative. Excellent."
- "Most beneficial of all. Really prepared me for giving the test. Ideas presented were not in the manual; taking the test taught me what to teach kids for taking the test."
- "Super, really good; we could have spent two days on it."

The procedures used to motivate students to try their best on tests were viewed less positively. For example:

- 53.3% of the teachers felt that the motivational procedures were difficult for students to understand.
- Only 38.3% of the teachers plan to use the motivational procedures in the future.
- Only 38.1% of the teachers felt that the procedures motivated students to improve their scores from practice test to practice test.

Although there were exceptions, typical teacher comments about the reinforcement procedures were as follows:

- "Don't know for sure how much kids got out of it."
- "Negative, because it was too time consuming; kids could not do it without my help."
- "Hard for slow students who did not get many points; more reinforcing for high students."
- "Good but not great. Too many kids and lots of cheating."
- "Not very reinforcing."

The teachers' ratings of the reinforcement procedures in conjunction with their comments indicate that the reinforcement procedures were the weakest part of the project.

In summary then, these data indicate that teachers generally felt very positive about the components of the project designed to teach test-taking skills to children and standardized test administration procedures to teachers. They were not as positive about the procedures designed to motivate students to try their best on tests. Teachers not only liked the project but felt that it was having a positive impact on students' test-taking skills and on the quality of test administration. A strong indicator of teachers' perceptions of the project's value is that they plan to continue using the materials in the future when there was no longer any "requirement" from the project to do so.

The following comments from teachers collected during the project debriefing underscore teachers' positive evaluation of the training materials and procedures.

- "Liked the program; really needed to teach these concepts; kids had good feelings."

- "Although it was hectic at the end, the training really showed during the test. This was the easiest administration of the SAT I've had in eight years. Really helped our English As A Second Language children who otherwise would have been wiped out by this experience. Wished we would have had similar training for math."
- "Enjoyed. Worthwhile. Kids were more relaxed."
- "Kids were very relaxed this year. Easiest administration of SAT in 10 years of teaching. Entire project easy to plan, prepare, and administer for teacher."
- "Excellent preparation for the test. Kids really learned to proofread and used these skills with other assignments."

Differences Between Groups on Outcome Variables

As noted earlier, classes participating in the project were randomly assigned to one of three groups: Experimental Group I (E1) received all components of the project including training students in test-taking skills (using filmstrips and practice tests), training teachers in standardized test administration procedures, and procedures to motivate students to do their best on standardized achievement tests. Participants in Experimental Group II (E2) received only the student training in test-taking skills (filmstrips and practice tests) and did not receive the teacher training in standardized test administration or the reinforcement procedures. It should be noted that although teachers in E2 were not explicitly trained in standardized test administration, the structured way in which practice tests were administered did provide them with frequent practice in administration of standardized tests which may have transferred to some degree to their administration of the actual standardized test. Because this training was implicit rather than explicit as it was in E1, it was not anticipated that it would have a very powerful effect on teachers' performance during the actual standardized test. Participants in the control group did not receive any of the project materials.

Table 40 summarizes the results for each of the outcome measures according to these three groups. Included in Table 40 are means, medians, control group standard deviations, probability levels from one-way analyses of variances, and measures of effect size differences between the groups for all participating students.

Results in Table 40 are also reported for three additional subgroups of students. First, only those students who received a majority of the treatment and eliminating students who were in special education programs or for whom English was not their primary language. As noted in the footnote to Table 40, this subgroup was defined as those students who viewed five or more of the filmstrips, took three or more of the practice tests, were not in special education programs or English As A Second Language programs, and had teachers who were not rated at the bottom of the scale on quality of implementation or support for the project. The second subgroup of students were those students who received all of the treatment. This category was defined in the same way as the preceding one except it was limited to those students who saw all nine filmstrips and took all seven practice tests. A third subgroup considered only students in Title I programs who received all of the treatment.

The analyses for these three additional subgroups were performed to determine if the program had differential effects for certain types of students. As can be seen in Table 40, the outcomes are very similar for all four groups. Therefore, the discussion which follows will refer to the total group of students except where noted. It should also be noted that the probability levels derived from the analyses of variance reported in the first two columns were computed based on mean differences between groups. As is well known, in cases where distributions of scores are skewed, medians rather

Table 40

Scores on Dependent Variables by Experimental Group

BEST COPY AVAILABLE

Variable		All Students	p[c]	Control Group SD	ES[d]	Students Receiving Majority of Treatment[a]	p[c]	Students Receiving All of Treatment[b]	Title I Students Receiving All of Treatment[b]
Teacher Attitude (Total)	$\bar{X}$	85.7 87.9 89.1	.000	12.3	.35	85.9 86.2 88.4	.014	83.7 85.9 87.0	*85.3 86.5 86.0
		C < 2 < 1				C < 1 < 2		1 < C < 2	1 < C < 2
	Md	85.5 87.2 89.8				85.5 86.7 87.4		81.3 85.5 91.6	85.4 85.3 87.5
Teacher Attitude (Opinion)	$\bar{X}$	14.7 14.9 15.1	.292	3.1	.10	13.9 15.0 15.6	.000	13.2 15.0 15.3	13.3 14.7 15.5
		1 < C < 2				1 < C < 2		1 < C < 2	1 < 2 < C
	Md	14.8 14.9 15.1				14.3 14.9 15.6		13.5 14.9 15.4	14.0 15.0 15.7
Teacher Attitude (Feeling)	$\bar{X}$	19.5 20.7 20.9	.000	3.1	.35	*19.6 20.7 20.1	.001	*19.6 20.6 20.1	19.4 20.3 20.7
		C < 1 < 2				C < 1 < 2		C < 2 < 1	C < 1 < 2
	Md	19.6 20.2 20.7				19.7 19.9 20.7		19.7 19.8 19.9	19.5 20.0 20.2
Teacher Attitude (Use)	$\bar{X}$	*24.8 25.9 25.6	.000	4.2	.55	*24.3 26.0 25.6	.000	23.1 26.0 25.7	*22.9 22.1 24.7
		1 < C < 2				1 < C < 2		1 < C < 2	1 < C < 2
	Md	23.3 25.8 25.6				23.1 25.9 26.9		21.6 25.9 26.8	21.4 24.6 26.2
Teacher Attitude (Increase)	$\bar{X}$	10.0 10.1 10.5	.000	2.5	.32	*10.1 10.0 10.3	.160	*10.1 9.7 10.1	9.4 10.0 10.7
		2 < C < 1				C < 2 < 1		1 < 2 < C	2 < C < 1
	Md	10.1 10.1 10.9				10.1 10.1 10.4		9.4 9.7 10.0	9.7 10.2 11.3
Teacher Attitude (Students' Feeling)	$\bar{X}$	15.3 16.2 18.3	.000	3.3	.82	15.3 16.6 17.6	.000	*15.3 17.2 15.8	15.6 16.5 18.1
		C < 2 < 1				C < 2 < 1		C < 1 < 2	C < 2 < 1
	Md	15.5 16.6 18.2				15.5 17.1 17.9		15.5 16.4 16.8	15.9 17.0 19.8
Teacher On-Task (Teacher-Directed)	$\bar{X}$	59.2 73.1 77.1	.000	49.0	.60	59.4 77.0 83.6	.000	59.4 71.5 79.7	55.0 74.5 78.2
		C < 2 < 1				C < 2 < 1		C < 2 < 1	C < 2 < 1
	Md	68.9% 89.1% 98.3%				68.8 95.4 99.0		68.8 95.8 96.0	68.9 96.1 97.1
Teacher On-Task (Student-Directed)	$\bar{X}$	*83.2 78.4 80.6	.048	37.7	.14	*83.1 81.2 81.6	.665	*83.1 81.2 79.7	*83.4 77.9 78.6
		C < 2 < 1				C < 2 < 1		C < 2 < 1	C < 2 < 1
	Md	87.7 88.7 92.8				88.0 88.7 93.8		88.0 88.9 93.8	87.8 88.0 94.3

[a] Eliminating students who saw less than 5 filmstrips, took less than 3 practice tests, had teachers who were rated low on quality of implementation or support, or were in special education programs, or had English as a second language.

[b] Eliminating students who saw less than 9 filmstrips, took less than 7 practice tests, had teachers who were rated low on quality of implementation or support, or were in special education programs, or had English as a second language.

[c] All probability estimates are based on one-way analyses of variance between means of the three groups. In many cases, distributions are substantially skewed so that medians are a better indicator of central tendency. Medians for each group on all variables are also reported. Asterisks are used to indicate where the order of groups differs depending on whether means or medians are reported. The order of groups represented in the chart always follows medians when there is a disagreement.

[d] The column labeled ES refers to the standardized mean differences between the highest and lowest group or $(\bar{X}_{high} - \bar{X}_{low}) \div SD_{control\ group}$. This measure has been recommended by Glass (1977) for examining the results of various studies using a common metric.

Table 40 (cont'd)

BEST COPY AVAILABLE

Variable		All Students					p[c]	Control Group SD	ES[d]	Students Receiving Majority of Treatment[a]					p[c]	Students Receiving All of Treatment[b]					Title I Students Receiving All of Treatment[b]				
Achievement Test (Student-Directed)	$\bar{X}$	*-.08		.08		.01				*.11		.28		.16		.08		.28		.33	-.86		-.40		-.13
		1	<	C	<	2	.033	.9	.19	1	<	C	<	2	.024	1	<	C	<	2	1	<	C	<	2
	Md	.19		.33		.38				.38		.42		.43		.38		.43		.48	-.85		-.28		.23
Achievement Test (Teacher-Directed)	$\bar{X}$	-.10		.04		.07				*.17		.12		.20		*.17		.12		.29	-.69		-.37		-.19
		1	<	C	<	2	.023	.9	.24	C	<	1	<	2	.508	C	<	1	<	2	1	<	C	<	2
	Md	.06		.09		.32				.22		.29		.40		.23		.23		.50	-.67		-.31		-.05
Achievement Test (Math)	$\bar{X}$	-.11		-.07		.19				.05		.06		.30		.13		.22		.30	-.65		-.41		-.11
		1	<	2	<	C	.000	.9	.36	2	<	1	<	C	.000	1	<	2	<	C	1	<	2	<	C
	Md	-.12		-.06		.24				.09		.11		.37		.13		.32		.79	-.62		-.48		-.18
Achievement Test (Total Reading)	$\bar{X}$	*-.10		.07		.05				*.11		.30		.20		.08		.26		.35	-.86		-.47		-.21
		1	<	C	<	2	.013	.9	.29	1	<	C	<	2	.058	1	<	C	<	2	1	<	C	<	2
	Md	.05		.22		.34				.22		.40		.40		.18		.40		.46	-.86		-.39		.08
Achievement Test (Total)	$\bar{X}$	-.12		.01		.13				.12		.17		.30		*.13		.34		.30	-.81		-.37		-.28
		1	<	2	<	C	.000	.9	.22	1	<	2	<	C	.014	1	<	2	<	C	1	<	C	<	2
	Md	.03		.13		.25				.21		.27		.37		.20		.36		.37	-.72		-.40		-.07
Quality of Test Administration	$\bar{X}$	*48.8		50.6		49.7				48.8		51.3		52.3		48.8		51.1		52.6	49.4		50.7		52.1
		C	<	2	<	1	.000	3.8	.28	C	<	2	<	1	.000	C	<	2	<	1	C	<	2	<	1
	Md	48.3		50.9		52.1				48.3		51.6		52.4		48.3		51.6		52.5	50.3		51.7		52.4
Student Attitude	$\bar{X}$	11.7		11.9		12.4				11.5		11.6		12.4		11.6		11.7		12.4	11.2		11.5		12.3
		2	<	1	<	C	.011	3.5	.20	1	<	2	<	C	.000	1	<	2	<	C	1	<	2	<	C
	Md	11.2		11.3		12.1				11.05		11.06		12.2		11.15		11.20		12.2	11.0		11.0		11.7
Student On-Task (Teacher-Directed)	$\bar{X}$	88.4		89.2		89.7				*89.4		89.5		89.2		*87.1		86.8		89.5	89.4		88.5		92.1
		1	<	C	<	2	.785	11.1	.32	1	<	C	<	2	.992	1	<	2	<	C	1	<	C	<	2
	Md	90.8		92.5		94.4				90.5		94.0		94.8		87.3		91.0		94.0	87.5		94.8		98.0
Student On-Task (Student-Directed)	$\bar{X}$	*90.5		90.6		89.9				89.5		90.9		93.0		*92.3		87.6		90.9	89.4		90.8		93.4
		C	<	1	<	2	.911	9.3	.14	2	<	C	<	1	.129	1	<	2	<	C	C	<	1	<	2
	Md	92.5		93.4		93.8				93.7		94.3		94.5		93.3		93.6		94.3	89.3		91.5		95.0

[a] Eliminating students who saw less than 5 filmstrips, took less than 3 practice tests, had teachers who were rated low on quality of implementation or support, or were in special education programs, or had English as a second language.

[b] Eliminating students who saw less than 9 filmstrips, took less than 7 practice tests, had teachers who were rated low on quality of implementation or support, or were in special education programs, or had English as a second language.

[c] All probability estimates are based on one-way analyses of variance between means of the three groups. In many cases, distributions are substantially skewed so that medians are a better indicator of central tendency. Medians for each group on all variables are also reported. Asterisks are used to indicate where the order of groups differs depending on whether means or medians are reported. The order of groups represented in the chart always follows medians when there is a disagreement.

[d] The column labeled ES refers to the standardized mean differences between the highest and lowest group or $(\bar{X}_{high} - \bar{X}_{low}) \div SD_{control\ group}$. This measure has been recommended by Glass (1977) for examining the results of various studies using a common metric.

than means are a better indicator of central tendency. In a few cases (noted by asterisks in the row for means), the median scores for groups were in a different order than means. In other cases, medians substantially reduced the differences or, in a few cases, increased differences between groups. Therefore, the probability levels given are only one source of information and should not be overinterpreted.

The most meaningful information about program effect is the Effect Size (ES) measure given as the last column for the "All Students" data. This Effect Size measure is an indicator of the standardized difference between the highest and lowest groups using the standard deviation of the control group as the standardizing metric. In most educational measures, an effect size of less than 1/3 of a standard deviation is not considered practically significant even though it may be statistically significant. Statistical significance indicates the probability of obtaining differences as large or larger as those observed in the experiment if one were to randomly draw samples of the same size from the same population. In cases where sample sizes are quite large (such as this project), it is not unusual to obtain statistical significance even though the differences are educationally and practically not very important. The effect size differences between groups reported in the last column of the "All Students" subcategory are computed based on medians instead of means and should be used in conjunction with probability levels and the order of groups in interpreting the results.

Additional information including sample sizes, means, standard deviations, and medians for all dependent variables broken down by experimental groups for each of the various subsamples is included in Appendix H. Also included in Appendix H are similar breakdowns for the different districts that participated in the project and descriptive statistics for each of the variables collected by the project. None of these more detailed data alter the basic interpretations to be presented in the following sections. The detailed data were not included in the main body of the report because the additional detail obscured rather than illuminated the major findings. Using the results reported in Table 40 as the basic information and supplementing it with other data as noted, the results of the project for each of the major dependent variables for each of the experimental groups are summarized below.

Teacher and student attitudes and behavior during standardized achievement tests. As can be seen in Table 40, there was approximately a third of a standard deviation difference on Teachers' Attitude Toward Standardized Achievement Tests between Experimental Group I and the Control Group, with Experimental Group II scoring in between. The major differences were in teachers' perceptions of the usefulness of standardized achievement tests and teachers' perceptions about how students felt about standardized achievement tests. The order of the groups on this outcome measure are what would have been predicted if the project were having its anticipated effect in terms of training teachers to be more competent and informed users of information from standardized achievement tests.

Teachers on-task behavior during the administration of the standardized achievement test was also improved. For both student- and teacher-directed subtests, there was approximately one-third of a standard deviation difference between Experimental Group I and the Control Group with Experimental Group II falling in between. The largest differences (.6 standard deviation units) was found on the teacher-directed subtest. This is not surprising since the procedures for teacher-directed subtests are much more complex in terms of proper test administration and require a higher level of skill and more constant involvement of the teacher. During the student-directed subtest, directions are given once at the beginning of the test and then students work on their own until the time limit has elapsed.

In interpreting the differences between groups for teacher on-task behavior, the definitions of on-task behavior described in the implementation section of this report should be kept in mind. As would be expected, teacher on-task time was defined similarly for both the instructional and data collection components. Therefore, conclusions about differences in on-task behavior of teachers indicate that teachers did indeed implement the types of things which were taught during the training. The real issue is, of course, whether these types of activities lead to a more valid test administration. Definitive answers to the question of whether these behaviors lead to more valid test scores are extremely complex.

The data indicating that teachers in Experimental Group I were more on-task during the administration of the standardized achievement test are supported by the ratings of quality of test administration. As noted in the implementation section, these ratings were done by observers who were not informed about the purpose of the experiment or the constituency of the groups. The differences between groups on quality of test administration are smaller (approximately a quarter of a standard deviation) but are in the direction which would be predicted if the project were having the anticipated effect. Again, the real meaning of these results can be interpreted best by looking at the type of items contained on the Quality of Test Administration Checklist described in the implementation section. Item level data from this checklist are contained in Appendix G. Taken together, the data from the teacher on-task behavior and the Quality of Test Administration Checklist indicate that the project had a positive effect on the procedures used by teachers to administer standardized achievement tests.

Data from the student on-task behavior during standardized achievement testing are more difficult to interpret. As shown in Table 40, there are virtually no differences--statistical or educational--between groups when the mean student on-task scores are considered. However, particularly for the teacher-directed subtest, these distributions are substantially skewed and medians indicate approximately a third of a standard deviation difference (favoring the control group for some subgroups and E2 for others). For the student-directed test, means and medians are in a different order but are all very close (approximately a tenth of a standard deviation difference between high and low). The flip-flopping of scores depending on which subtest, which subgroup, and whether means or medians are examined suggests that differences are not educationally meaningful. Furthermore, the very high levels of student on-task behavior across all subtests and subgroupings (87% - 98%), and the fact that student on-task and achievement test scores using the student as the unit of analysis is uncorrelated (r ranges from -.00 to .01; see Table 43) suggests that the measure of student on-task behavior may not have been an accurate measure of on-task behavior. There is an extensive body of literature which suggests that on-task behavior is moderately related to achievement levels in instructional settings, and it is reasonable to assume that on-task behavior should be correlated with scores on standardized achievement tests. Even though the interrater consistency for the student on-task behavior measures were high, the data reported above raise concerns as to whether the essence of student on-task behavior during standardized testing was really measured.

Student achievement. As can be seen from Table 40, when median scores are considered, Group II had the highest standardized achievement test scores for all of the reading subtests; and the Control Group had the highest scores for the math subtest and for the total test battery. Statistical tests of significance computed from the analysis of variance using means were significant for all achievement test measures, sometimes favoring Group II and sometimes favoring the Control Group. The magnitude of the differences between groups is generally increased slightly when medians instead of means are used, but the order of differences changes for the total reading score and the student-directed reading test score. Although statistically significant, the differences are generally small (average ES = .26, ranging from .19 to .36 of a standard deviation difference). These already small differences are reduced even further when the analyses are limited to those students who received the majority or all of the treatment and only reading subtests are considered (average ES = .19). In this subset of the data, E2 had the highest scores for all three reading tests with E1 receiving second highest for one of the subtests, and the Control group for the other two subtests.

Math subtest scores were collected and analyzed for two reasons. First, if the treatment were effective, it would be interesting to see if the results generalized to other testing areas for which no explicit training was included. Secondly, if the treatment did not appear to be effective, math scores could be used as a partial way of checking the comparability of the groups. In other words, if the scores between the groups on math were radically different, one would be concerned about the comparability of the groups before treatment began because one would not expect the treatment to have as powerful an effect on the math subscores as on the reading subscores. The fact that the math and total subtest scores are the only subtests where the control group scored the highest, and the fact that some of these effect sizes are substantial (e.g., .66 in the "all of the treatment" subgroup) suggests that there may be some sample comparability problems. This issue is discussed further in a subsequent section. Although many of the differences between groups are statistically significant, they are relatively small. The only consistent finding in the order of these differences is that Group I was regularly the lowest-scoring group. For reading test scores, Group II was always the highest-scoring group, while for the math and total scores, the Control Group was the highest-scoring group. The fact that the Control Group scored highest on the total test is at least partly a function of the heavy contribution of math scores to this total battery score.

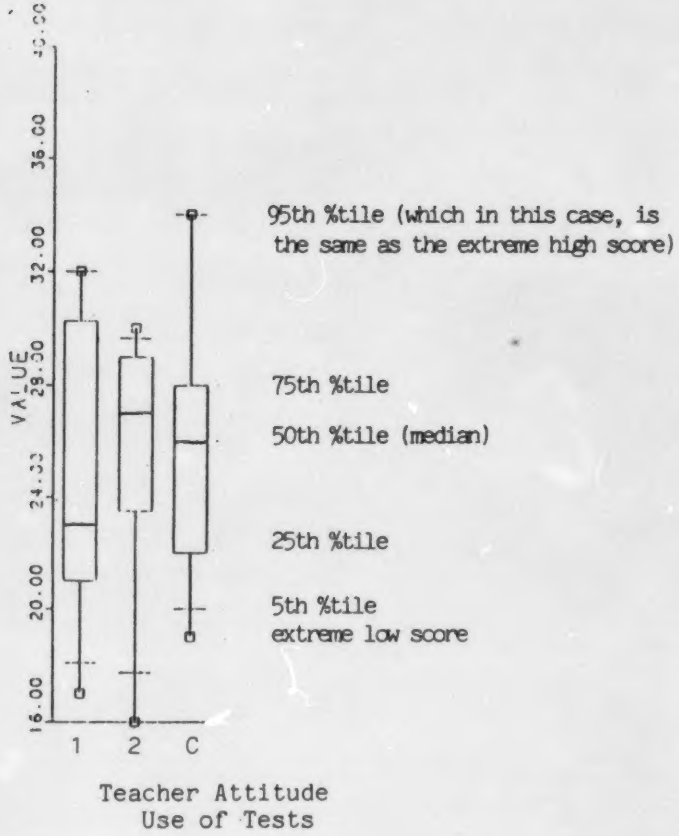
Although not particularly clear-cut, these data indicate that the project did not have a meaningful effect on student achievement test scores as had been hypothesized. If the project was contributing to standardized achievement tests being a more valid indicator of students' knowledge in a particular content area, it was anticipated that scores would increase. In other words, previous research had indicated that students' scores on standardized achievement tests were at least partly influenced by the format of the test, students' test-taking skills, and the degree to which students were motivated to do well on the test. Each of these confounding variables seemed to result in students appearing to know less than they really did. Therefore, it was hypothesized that the experimental procedures would remove these influences causing test scores to increase. This clearly did not happen.

Of course, it is possible for the test scores to become more valid even if the scores do not increase. This could happen if the pattern of correct answers within a subtest changed. However, this is a much more unlikely occurrence and the most reasonable conclusion from the data is that the intervention procedures had little, if any, effect on students' scores on standardized tests.

Another possible explanation for the observed standardized achievement test scores is that the reinforcement procedures (which was that portion of the program which was least well received by teachers) actually depressed students' scores instead of the anticipated elevating effect. This hypothesis is given some support from the fact that Group II which did not receive the reinforcement procedures was consistently the highest-scoring group on those measures which were most directly related to the treatment. However, even here, the differences between E2 and the Control group are always small (ranging from .05 to .23 when all students are considered, and .00 to .27 when only students who received the majority or all of the treatment are considered). The average differences (.04 to .13) are not large enough to be practically significant, and given the fact that scores flip-flop from group to group and from test to test suggests that random fluctuation is a more plausible explanation.

The lack of effect from the intervention program on students' standardized achievement test score is shown graphically in Figure 5 which contains Box and Whisker diagrams for each of the achievement subtests (modified from Tukey, 1977). As shown on the next page, the box of a Box and Whisker diagram depicts the interquartile range of scores. The "whisker" extending from the box shows the range of scores with the crosshatch on the "whisker" showing the 5th or the 95th percentile and the small box at the end of the "whisker" showing the most extreme scores.

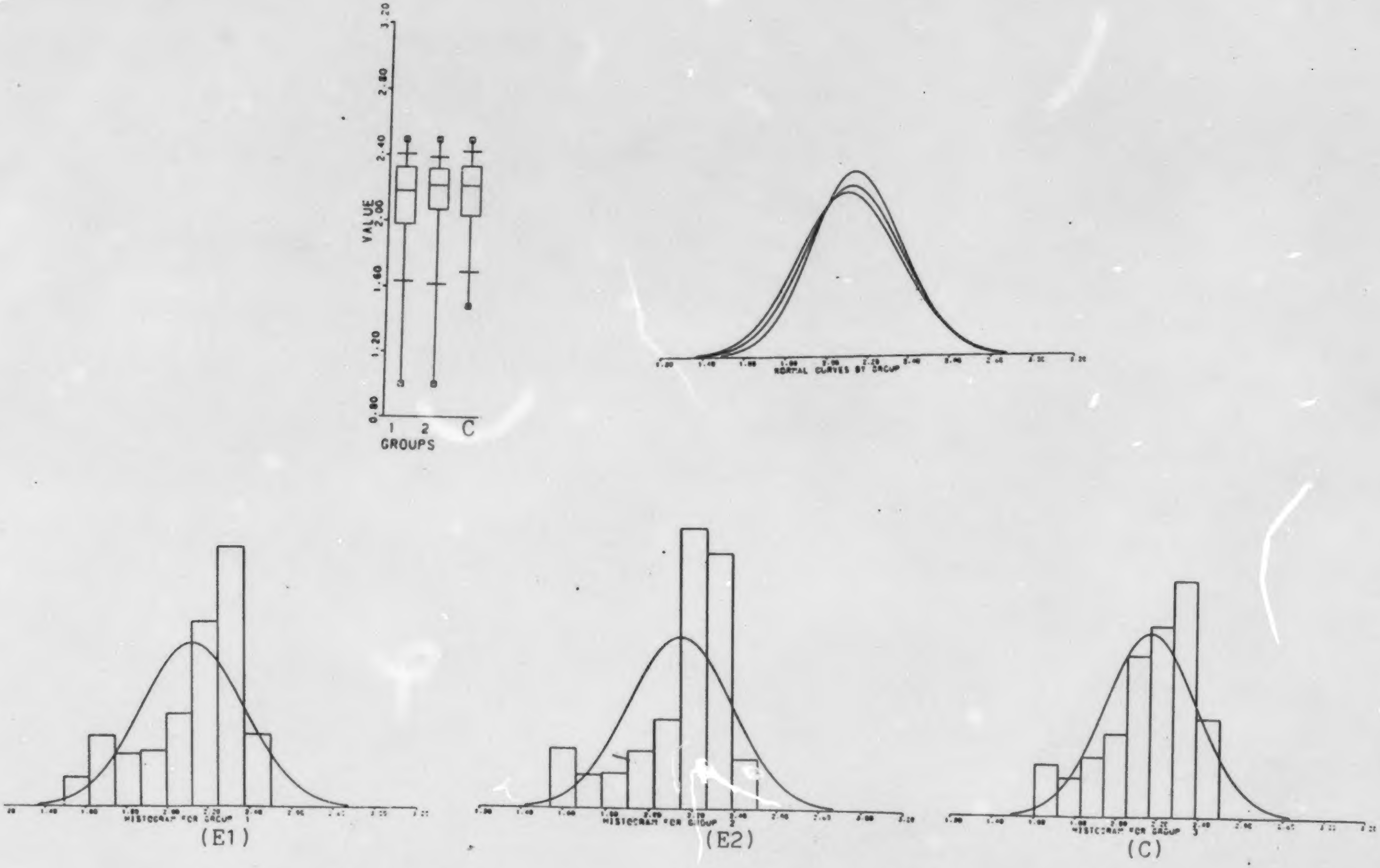
As can be seen on the Box and Whisker diagram for the student-directed reading subtest of the achievement test scores in Figure 5, the interquartile ranges for all three groups are almost completely overlapping (the same degree of overlap is present for virtually all of the dependent variables). The normal curves to the right of the Box and Whisker diagram for this subtest shows the amount of overlap which would occur if normal curves were constructed using the mean and standard deviation from the different groups.



As would be expected, overlap is again almost complete. The normal curves with overlaying bar graphs shown below the Box and Whisker diagrams show the actual data distribution (bar graphs) and how well they conform to normal curves. As can be seen, for all three groups the data are negatively skewed; worse so for Groups I and II. This negative skewing is also apparent from the longer tails on the lower portion of the Box and Whisker diagrams for Groups I and II. Figure 5 shows the same type of patterns for all of the achievement test scores.* The basic message from these diagrams is similar to the conclusion outlined above, i.e., differences between the three groups on achievement test scores are small and educationally insignificant.

Figure 5

Box and Whisker Diagrams and Normal Curve Representations
for Student-Directed Reading Achievement Subtest
(Square Root Transformation)



(E1)

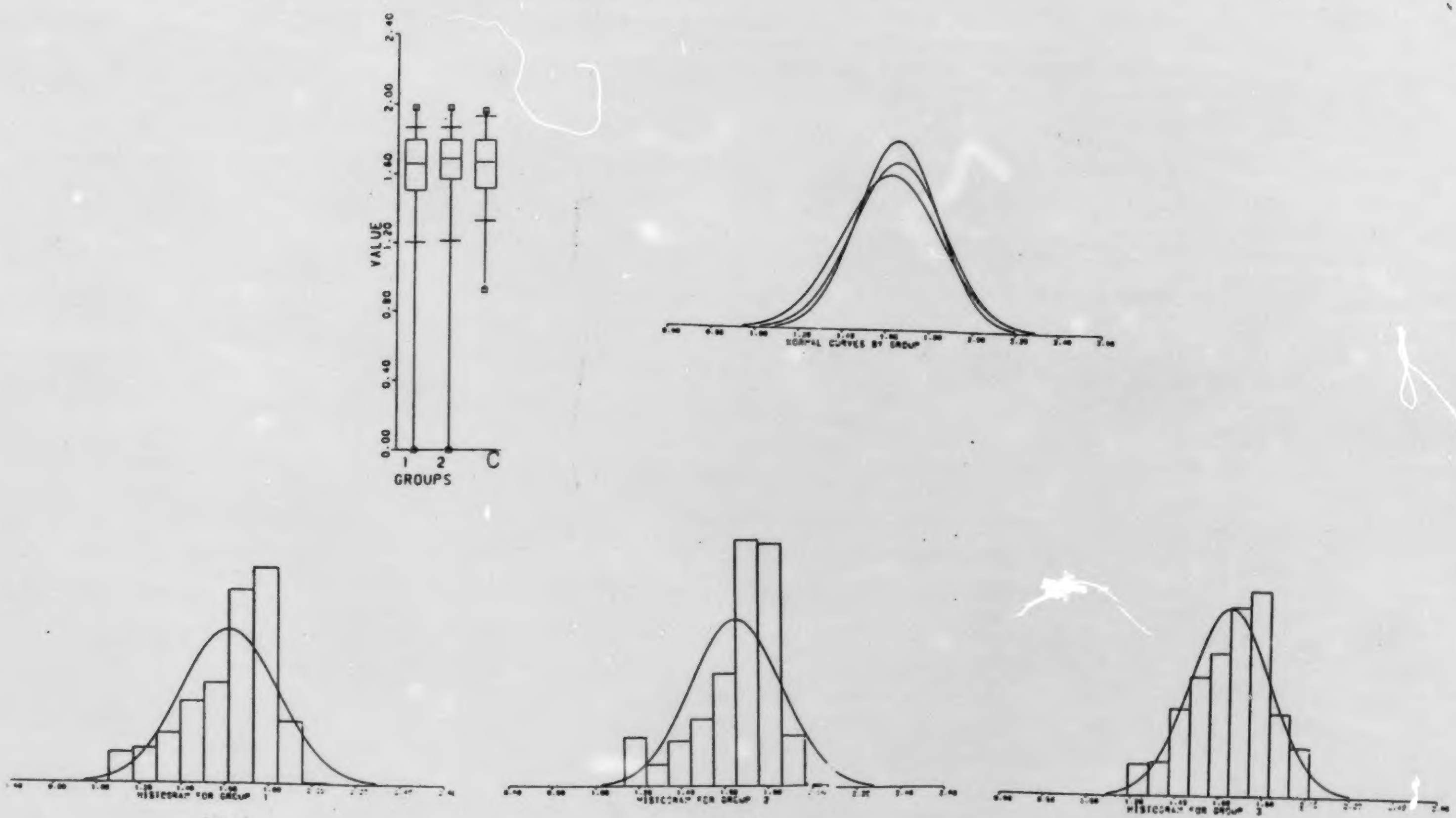
(E2)

(C)

Note: Data used in the Box and Whisker diagrams have been transformed as indicated using square root or log transformations to make the distributions of each group more comparable (Tukey, 1977).

Figure 5 (cont'd)

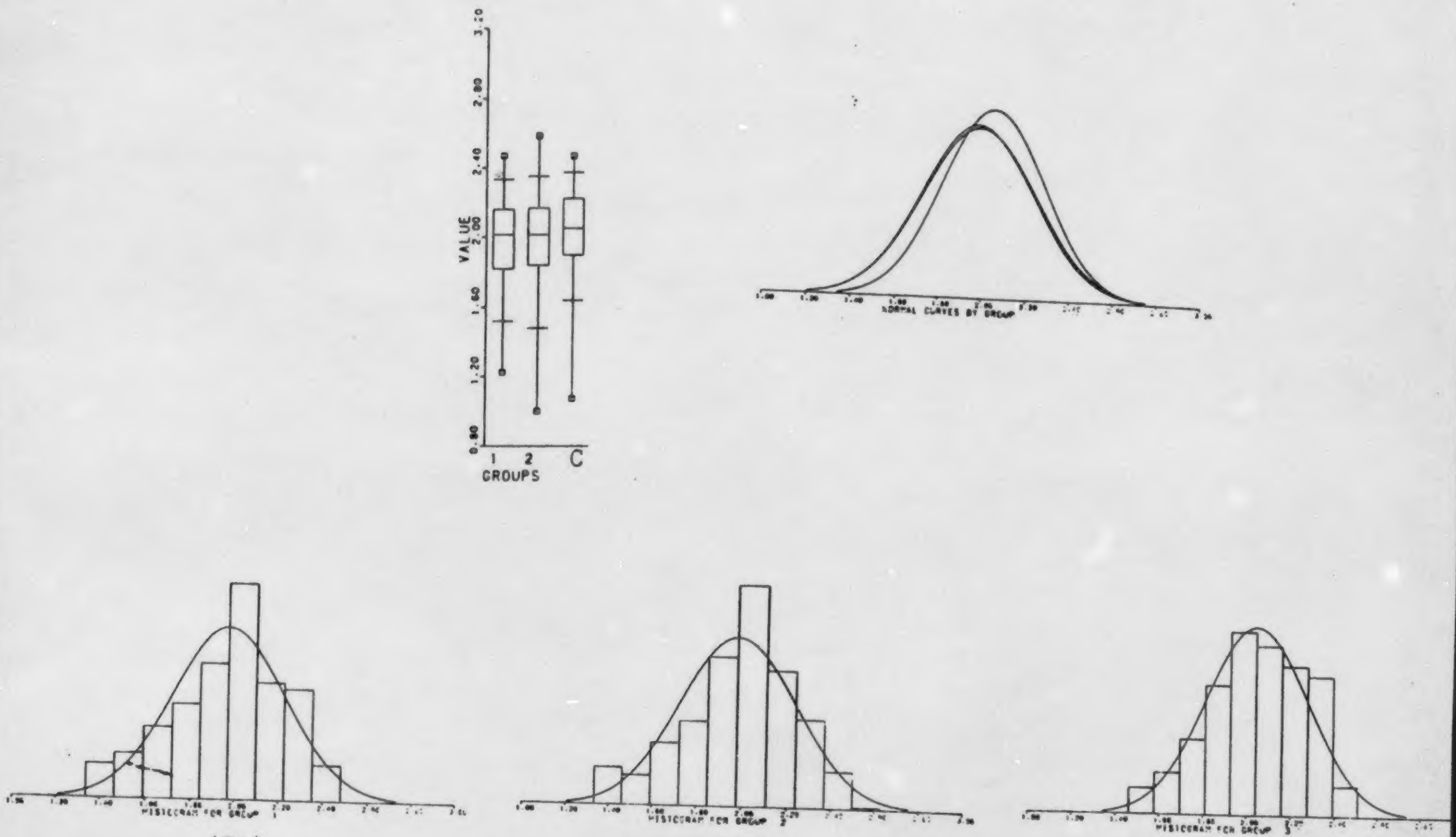
Box and Whisker Diagrams and Normal Curve Representations
for Teacher-Directed Reading Subtest
(Log Transformation)



(E1) (E2) (C)

Figure 5 (cont'd)

Box and Whisker Diagrams and Normal Curve Representations
for Total Math Subtests
(Square Root Transformation)



(E1)

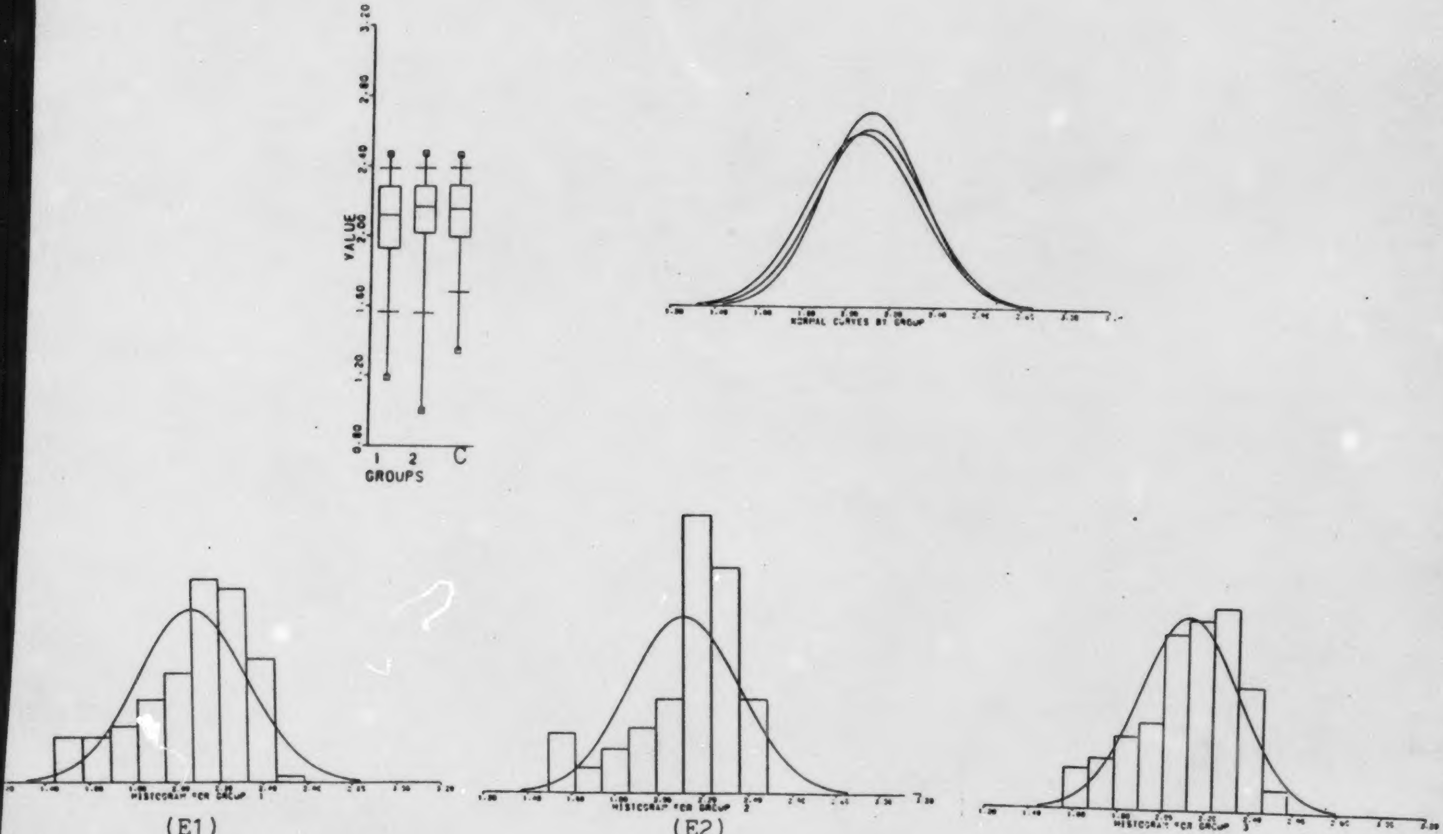
(E2)

(C)

BEST COPY AVAILABLE

Figure 5 (cont'd)

Box and Whisker Diagrams and Normal Curve Representations
for Total Reading Subtests
(Square Root Transformation)



(E1)

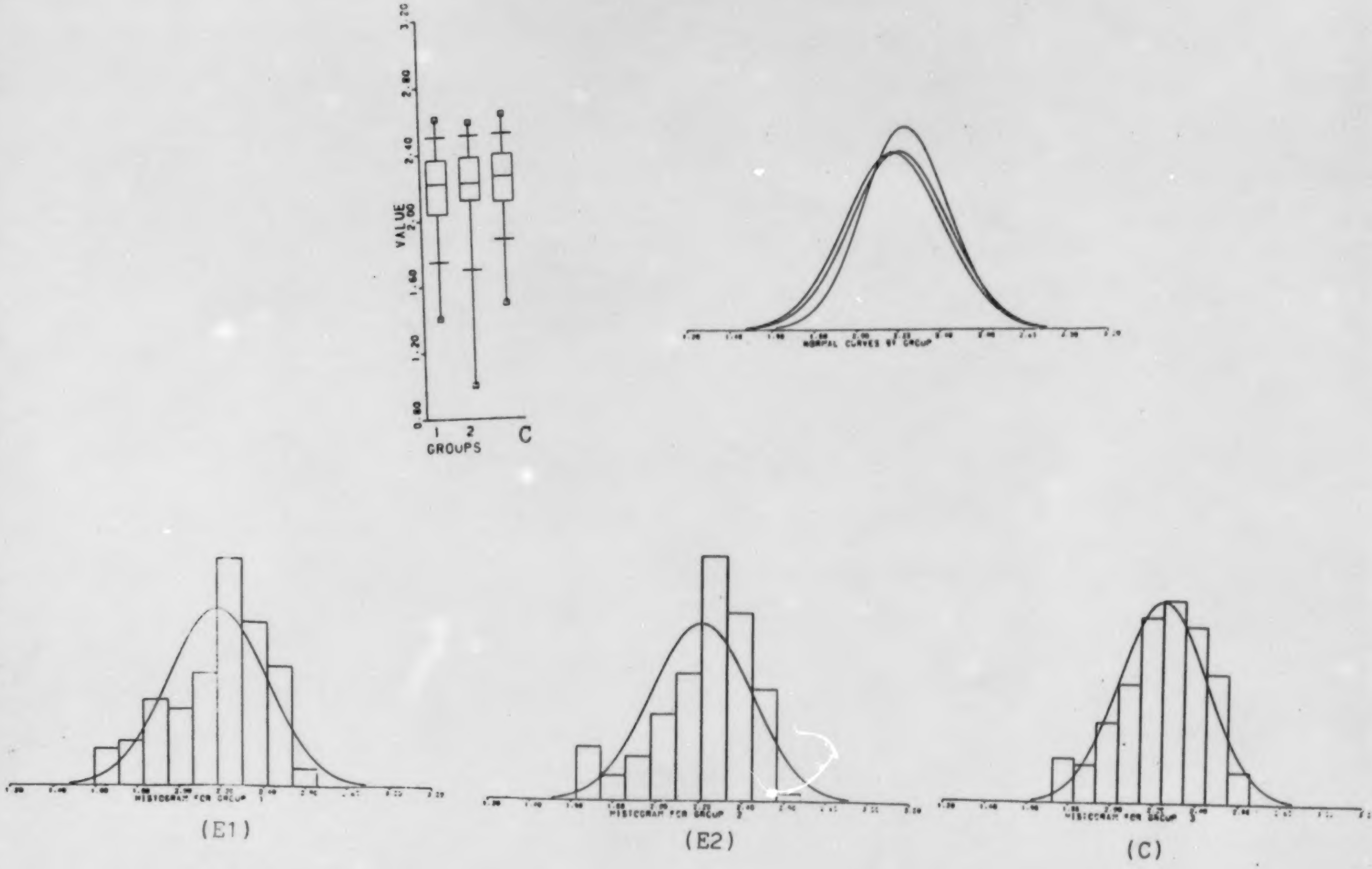
(E2)

(C)

BEST COPY AVAILABLE

Figure 5 (cont'd)

Box and Whisker Diagrams and Normal Curve Representations
for Total Achievement Test Scores
(Square Root Transformation)



(E1)

(E2)

(C)

BEST COPY AVAILABLE

Several other analyses were done to better understand the pattern of scores on standardized achievement tests. Crosstabs with various other demographic characteristics and implementation variables were computed with total achievement test scores. These analyses are reported in Table 41. As can be seen, test scores were substantially correlated with the number of filmstrips viewed by students, number of practice tests taken, teacher support of the program, quality of teacher implementation, the number of reinforcement points earned, the mean percentage correct on practice tests, and whether students were in special education, Title I, or English As A Second Language programs. Some of this relationship was curvilinear and consequently does not appear in the correlational data reported in Table 42 later in this report.

The fact that students who saw the most filmstrips received the highest scores on achievement tests could be viewed as an indicator that the program had indeed contributed to higher scores on standardized achievement tests. However, this type of correlational data is a much weaker indicator of causality than the data reported previously from Table 40 because a number of directionality and third variable explanations are plausible explanations (even though they are impossible to examine directly from the available data). For example, in E1 the average achievement test score (reported in Z scores) went from -.69 for students viewing 1 to 6 filmstrips, to -.51 for students viewing 7, to -.14 for students viewing 8, to -.03 for students viewing 9 filmstrips--an increase of .66 standard deviation units. However, those students who saw all 9 filmstrips are more likely to be students who have good attendance records in school and consequently are being exposed to more instruction, probably have better attitudes towards school as indicated by their better attendance; and likely come from homes where more value is placed on education. Although most teachers did make substantial efforts to

Table 41

Total Achievement Test Scores by Group by Various Independent Measures

No. of Filmstrips Viewed	Experimental Group		
	I	II	
1-6	n = 30 (5.9%) $\bar{X}$ = -.69 SD = 1.23	n = 20 (4.8%) $\bar{X}$ = -.22 SD = 1.07	p < .06
7	n = 31 (6.0%) $\bar{X}$ = -.51 SD = 1.25	n = 58 (14.0%) $\bar{X}$ = .05 SD = 1.12	
8	n = 119 (23.5%) $\bar{X}$ = -.14 SD = 1.00	n = 99 (24.0%) $\bar{X}$ = -.17 SD = 1.05	
9	n = 327 (65.0%) $\bar{X}$ = -.03 SD = .98	n = 240 (58.0%) $\bar{X}$ = .09 SD = .98	
	n = 507 $\bar{X}$ = -.12 SD = 1.04	n = 417 $\bar{X}$ = .008 SD = 1.02	

No. of Practice Tests Taken	Experimental Group		
	I	II	
1-5	n = 62 (12.0%) $\bar{X}$ = -.32 SD = 1.10	n = 40 (9.6%) $\bar{X}$ = -.32 SD = 1.03	p < .05
6	n = 141 (28.0%) $\bar{X}$ = -.17 SD = 1.03	n = 101 (24.0%) $\bar{X}$ = -.16 SD = 1.09	
7	n = 304 (60.0%) $\bar{X}$ = -.05 SD = 1.02	n = 274 (66.0%) $\bar{X}$ = .13 SD = .98	
	n = 504 $\bar{X}$ = -.12 SD = 1.04	n = 416 $\bar{X}$ = .01 SD = 1.02	

Teacher Support of Program	Experimental Group		
	I	II	
1	n = 24 (4.7%) $\bar{X}$ = -.70 SD = .98	n = 24 (6.0%) $\bar{X}$ = -.12 SD = .93	p < .05
2	n = 173 (34.0%) $\bar{X}$ = -.14 SD = .98	n = 178 (43.0%) $\bar{X}$ = -.10 SD = 1.04	
3	n = 312 (61.3%) $\bar{X}$ = -.07 SD = 1.06	n = 215 (52.0%) $\bar{X}$ = .11 SD = 1.01	
	n = 509 $\bar{X}$ = -.12 SD = 1.04	n = 417 $\bar{X}$ = .008 SD = 1.02	

Quality of Teacher's Implementation	Experimental Group		
	I	II	
1	n = 95 (18.7%) $\bar{X}$ = -.36 SD = 1.03	n = 50 (12.0%) $\bar{X}$ = 0.003 SD = 1.04	p < .05
2	n = 99 (19.4%) $\bar{X}$ = -.22 SD = 1.07	n = 154 (37.0%) $\bar{X}$ = -.21 SD = 1.08	
3	n = 315 (62.0%) $\bar{X}$ = -.02 SD = 1.02	n = 213 (51.0%) $\bar{X}$ = .17 SD = .95	
	n = 509 $\bar{X}$ = -.12 SD = 1.04	n = 417 $\bar{X}$ = .008 SD = 1.02	

BEST COPY AVAILABLE

Table 41 (cont'd)

$\bar{X}$ No. of Reinforcement Points Earned Per Test	Experimental Group I
1-2.5	n = 95 (19.0%) $\bar{X}$ = -.69 SD = .80
2.51-3.5	n = 133 (27.0%) $\bar{X}$ = -.35 SD = .97
3.51-4.5	n = 117 (24.0%) $\bar{X}$ = .08 SD = .99
4.51-5.5	n = 93 (19.0%) $\bar{X}$ = .30 SD = .96
5.51-9.5	n = 49 (9.9%) $\bar{X}$ = .57 SD = .99
	n = 492 $\bar{X}$ = -.10 SD = 1.03 r_{xy} = .37

Mean % Correct on Practice Tests	Experimental Group I	Experimental Group II
1-60%	n = 20 (4.1%) $\bar{X}$ = -1.67 SD = .82	n = 54 (13.0%) $\bar{X}$ = -1 16 SD = 1.25
60.1-80%	n = 135 (27.0%) $\bar{X}$ = -.98 SD = .78	n = 90 (22.0%) $\bar{X}$ = -.65 SD = .79
80.1-85%	n = 75 (15.0%) $\bar{X}$ = -.34 SD = .65	n = 37 (9.0%) $\bar{X}$ = -.11 SD = .60
85.1-90%	n = 81 (16.5%) $\bar{X}$ = .07 SD = .68	n = 57 (14.0%) $\bar{X}$ = .21 SD = .58
90.1-95%	n = 119 (24.0%) $\bar{X}$ = .67 SD = .58	n = 94 (23.0%) $\bar{X}$ = .55 SD = .53
95.1-100%	n = 62 (13.0%) $\bar{X}$ = .87 SD = .68	n = 81 (20.0%) $\bar{X}$ = .79 SD = .75
	n = 492 $\bar{X}$ = -.10 SD = 1.03	n = 413 $\bar{X}$ = .01 SD = 1.03

$p < .10$

Student in Title I?	Experimental Group I	Experimental Group II	Control
No	n = 353 (69.0%) $\bar{X}$ = .31 SD = .82	n = 295 (71.0%) $\bar{X}$ = .23 SD = .98	n = 349 (77.0%) $\bar{X}$ = .34 SD = .85
Yes	n = 155 (31.0%) $\bar{X}$ = -1.11 SD = .77	n = 120 (29.0%) $\bar{X}$ = -.52 SD = .95	n = 102 (23.0%) $\bar{X}$ = -.60 SD = .78
	n = 508 $\bar{X}$ = -.12 SD = 1.03	n = 415 $\bar{X}$ = .009 SD = 1.02	n = 451 $\bar{X}$ = .13 SD = .92

$p < .0004$

BEST COPY AVAILABLE

Table 41 (cont'd)

Student in Special Education?	Experimental Group I	Experimental Group II	Control	
No	n = 474 (93.0%) $\bar{X}$ = -.02 SD = .98	n = 382 (92.0%) $\bar{X}$ = .11 SD = .94	n = 394 (87.0%) $\bar{X}$ = .29 SD = .81	p < .0004
Yes	n = 34 (6.7%) $\bar{X}$ = -1.58 SD = .57	n = 33 (8.0%) $\bar{X}$ = -1.18 SD = 1.22	n = 54 (13.0%) $\bar{X}$ = -.99 SD = .82	
	n = 508 $\bar{X}$ = -.12 SD = 1.04	n = 415 $\bar{X}$ = .009 SD = 1.02	n = 451 $\bar{X}$ = .13 SD = .92	

Student With English as a Second Language?	Experimental Group I	Experimental Group I	Control	
No	n = 492 (97.0%) $\bar{X}$ = -.08 SD = 1.01	n = 401 (97.0%) $\bar{X}$ = .04 SD = 1.01	n = 442 (98.0%) $\bar{X}$ = .14 SD = .91	p < .0004
Yes	n = 16 (3.2%) $\bar{X}$ = -1.38 SD = 1.09	n = 14 (3.4%) $\bar{X}$ = -.88 SD = 1.04	n = 9 (2.0%) $\bar{X}$ = -.50 SD = 1.10	
	n = 508 $\bar{X}$ = -.12 SD = 1.04	n = 415 $\bar{X}$ = .008 SD = 1.02	n = 451 $\bar{X}$ = .13 SD = .92	

Teacher's Evaluation of Project	Experimental Group I	Experimental Group II	
1-1.5	n = 98 (19.0%) $\bar{X}$ = -.12 SD = .99	n = 117 (28.0%) $\bar{X}$ = -.08 SD = 1.03	
1.6-2.5	n = 310 (61.0%) $\bar{X}$ = -.16 SD = 1.05	n = 247 (59.0%) $\bar{X}$ = .01 SD = 1.06	p < .05
2.6-3.5	n = 101 (19.8%) $\bar{X}$ = -.01 SD = 1.04	n = 53 (12.7%) $\bar{X}$ = .19 SD = .79	
	n = 509 $\bar{X}$ = -.12 SD = 1.04	n = 417 $\bar{X}$ = .008 SD = 1.02	

BEST COPY AVAILABLE

do make-ups with the filmstrips, the fact that many students did not see some of the filmstrips makes these explanations plausible and makes it somewhat unreasonable to suppose that the viewing of filmstrips resulted in higher test scores alone. The same explanations can be offered for each of the other variables where it looks like increased exposure or participation in the program resulted in higher test scores.

Scores for different groups of students (Title I, Special Education, and English As A Second Language) are exactly what one would expect to see based on our knowledge of the types of children who usually participate in those programs. Achievement test scores for these variables and their predicted direction lends some credibility to the test being used and hence, more confidence to the results reported above concerning the achievement test.

Table 42 contains the intercorrelation matrix for the principal variables, both dependent, independent, and descriptive included in the project. Generally, the correlations reported in this table support the kinds of findings reported above. For example, student on-task behavior, teacher on-task behavior, teacher attitude and student attitude, and quality of test administration are generally uncorrelated with achievement test scores. The consistency of these findings lends further support to the conclusion that the procedures as implemented had very little impact on achievement test scores.

Conclusions

The basic purpose of this project was to develop, implement, and evaluate training materials and procedures which would result in more valid standardized achievement test scores as a result of improvements in (a) students' test-taking skills, attitudes, and motivation towards test taking,

Table 42

Intercorrelation Matrix for Project Variables

	Teacher Attitude (Opinion about Tests)	Teacher Attitude (Feeling, Abt. Admin.)	Teacher Attitude (Usefulness of Tests)	Teacher Attitude (Should Use Be Incr'd)	Teacher Attitude (Student's Feelings)	Teacher On-Task TD	Teacher On-Task SD	Student On-Task TD	Student On-Task SD	Achievement Test SD	Achievement Test TD	Achievement Test Math	Achievement Test Reading	Achievement Test Total	Quality of Test Administration	# of Filmstrips Viewed	# of Practice Tests Taken	# of Reinforcement Points Earned	Average Correct on Practice Test	Student in Title I?	Student in Special Education?	English as a Second Language?	Teacher Support	Quality of Implementation	Teacher Evaluation (Total)	Teacher Evaluation (Filmstrips)	Teacher Evaluation (Practice Tests)	Teacher Evaluation (Proj. Communication)	Teacher Evaluation (Data Collection Act.)	Teacher Evaluation (General)	Teacher Evaluation (Reinforcement)	Teacher Evaluation (Spring Workshop)	Test-Wiseness	Test-Taking Skills (Deductive)	Test-Taking Skills	Student Attitude Towards Testing
Teacher Attitude (Opinion about Tests)		.42	.55	.31	.34	-.002	.09	.08	-.20	.03	.002	.02	.04	.04	.06	-.04	-.03	.10	.03	-.04	-.01	-.03	.09	-.10	-.17	-.02	-.32	.31	-.27	-.03	-.14	-.16	.02	.05	.03	.00
Teacher Attitude (Feeling about Administration)			.44	.32	.48	.09	.11	.04	-.008	.05	.05	.003	.06	.03	.03	-.07	.01	-.03	.15	.01	-.05	-.01	-.30	-.21	-.21	[illegible]	-.15	.01	-.16	-.15	-.34	-.28	.05	-.01	-.02	-.01
Teacher Attitude (Usefulness of Tests)				.49	.18	.08	.17	.14	-.06	.05	.005	-.004	.04	.02	.02	-.003	-.03	.04	-.04	-.10	.02	-.05	-.14	-.14	-.10	.04	-.10	.18	-.08	-.12	-.03	-.21	.05	.01	.00	.05
Teacher Attitude (Should Use be Increased)					.18	.08	-.06	.11	.04	-.06	-.10	-.07	-.09	-.10	-.10	.01	-.01	.17	-.14	.03	.05	.02	-.30	-.11	.02	.06	.06	-.10	-.02	-.04	.02	.03	-.07	-.08	-.06	.07
Teacher Attitude (Students' Feelings)						-.11	.08	.07	.20	-.03	.03	.03	-.005	-.003	-.22	-.04	-.04	-.01	.05	.05	-.06	.01	-.10	.10	-.16	-.04	-.06	-.12	.01	-.12	-.53	-.61	-.02	-.00	-.01	-.04
Teacher On-Task TD							.48	-.009	.009	-.04	-.05	-.09	-.06	-.07	.35	-.05	-.08	-.02	-.10	-.10	.01	-.001	.05	.15	.15	.41	.09	.21	-.12	-.06	.29	.28	.02	.03	.00	[illegible]
Teacher On-Task SD								.02	-.04	-.02	-.005	.06	-.02	.03	.30	-.05	-.04	.11	-.05	-.07	.04	.02	.29	.13	.09	.12	-.09	.11	-.19	-.10	-.13	-.02	.04	-.01	.06	[illegible]
Student On-Task TD									.25	-.009	.01	.001	.007	.007	.22	-.04	-.15	.10	-.10	.01	-.03	.09	.05	.12	-.19	-.07	-.21	-.08	-.15	-.26	-.18	.05	.03	-.01	.10	-.02
Student On-Task SD										.04	.06	.07	.009	.03	.10	-.10	-.17	-.25	-.06	-.03	-.04	.06	-.03	.32	.12	.19	.22	-.23	.15	-.04	.26	-.18	.07	.01	.17	.05
Achievement Test SD											.76	.61	.94	.87	.07	.10	.07	.35	.65	-.50	-.39	-.17	.08	.09	.05	.07	-.01	.10	.00	.04	.05	.07	.57	.40	.34	-.04
Achievement Test TD												.61	.87	.83	.08	.08	.07	.35	.62	-.43	-.31	-.14	.03	.12	.03	.07	-.01	.08	-.00	.03	.02	-.00	.53	.36	.36	-.04
Achievement Test (Math)													.67	.85	.05	.10	.13	.31	.55	-.37	-.28	-.09	.10	.19	-.03	-.01	-.06	.00	-.05	.00	-.04	.04	.53	.27	.39	-.03
Achievement Test (Reading)														.93	.05	.10	.09	.39	.69	-.48	-.39	-.18	.05	.07	.04	.05	-.01	.11	-.01	.04	.04	.05	.61	.40	.38	-.04
Achievement Test (Total)															.09	.11	.12	.37	.67	-.48	-.38	-.17	.09	.13	.02	.05	-.05	.09	-.03	.02	.00	.07	.64	.38	.42	-.04
Quality of Test Administration																-.03	-.02	-.00	.07	-.11	.02	.03	.46	.45	-.10	.11	-.30	.11	-.34	-.15	.00	.26	.12	.12	.10	-.08
# of Filmstrips Viewed																	.80	-.08	.02	.02	-.04	-.00	.06	.02	.07	.04	.07	.01	.08	.05	.02	.02	.08	.01	.08	.03
# of Practice Tests Taken																		-.14	.02	.01	-.06	-.01	.02	-.01	.04	.05	-.01	-.05	.08	.03	-.01	.12	.06	.02	.12	.01
# of Reinforcement Points Earned																			.38	-.18	-.12	.10	-.05	.01	.02	-.00	-.02	.06	-.15	.06	-.07	.16	.33	.06	.13	-.03
Average Correct on Practice Test																				-.25	-.28	-.14	.03	.01	-.05	-.06	-.10	-.03	-.34	.12	-.06	.04	.46	.31	.28	-.07
Student in Title I?																					.21	.13	-.01	.03	-.05	-.11	-.06	-.11	-.01	-.01	-.11	.02	-.37	-.29	-.26	-.01
Student in Special Education?																						.00	-.02	.02	-.02	-.04	-.01	-.02	-.05	-.01	.09	-.02	-.26	-.31	-.22	.06
English as a Second Language?																							.01	-.01	-.03	-.02	-.01	-.03	-.06	-.04	-.06	.06	-.02	-.07	-.12	-.00
Teacher Support																								.52	.26	-.36	-.33	-.05	.14	-.36	-.25	-.10	.06	.08	.09	-.11
Quality of Implementation																									-.06	-.02	-.13	-.19	.15	-.23	.01	-.00	.09	.07	.11	-.08
Teacher Evaluation (Total)																										.80	.83	.35	.43	.86	.81	.62	.05	.01	-.01	.15
Teacher Evaluation (Filmstrips)																											.64	.39	.25	.60	.64	.51	.11	.01	.23	.11
Teacher Evaluation (Practice Tests)																												.27	.42	.65	.54	.16	.00	-.03	-.06	.13
Teacher Evaluation Project Communication																													.12	.25	.03	.13	.08	.05	.05	.07
Teacher Evaluation Data Collection Activ.																														.14	[illegible]	.04	-.02	[illegible]	.02	.08
Teacher Evaluation (General)																															.71	.49	.03	.03	[illegible]	.15
Teacher Evaluation (Reinforcement)																																.44	.03	.10	.00	.18
Teacher Evaluation (Spring Workshop)																																	.12	.00	.05	.05
Test-Wiseness																																		.29	.30	-.03
Test-Taking Skills (Deductive)																																			.23	-.10
Test-Taking Skills																																				-.05
Student Attitude Towards Testing																																				

and (b) teachers' attitudes towards standardized tests and quality of test administration. As measured by the project developed instruments, the intervention procedures did result in improved teachers' attitude towards tests and quality of test administration. Furthermore, teachers were enthusiastically supportive of the materials, plan to continue using them, and felt that the materials resulted in substantial improvement in students' test-taking abilities and students' attitude towards tests. However, the more objective data collected by the project indicate that there were no increases in students' test-taking skills, attitudes toward standardized testing, or performance on standardized achievement tests.

These data raise some perplexing questions, both in view of previous research and in view of teachers' perceptions about the effectiveness of the project. First, as reported in the review of literature in Chapter II, previous research (from both published and unpublished sources) indicates that training students in test-taking skills or providing them with practice in taking tests has a substantial effect (approximately 2/3 of a standard deviation) on test scores. Even when the results of that research are limited to high-quality studies, the average effect attributable to training was approximately a third of a standard deviation. The intervention in this project combined both training in test-taking skills and practice on tests similar to the standardized achievement tests the students would be taking. In addition, previous research reported in the review of literature has also suggested that areas such as checking work, systematic elimination, problem attack strategies, reduction of test anxiety, examiner/examinee relationships, advance notification, and feedback on test performance are all positively correlated with scores on standardized achievement tests. All of these factors were included in the training packages designed. Finally, when compared to the interventions in previous research, the training delivered to these students was a relatively intense, systematically delivered training experience of long duration, with strong follow-up and monitoring.[1] In spite of this, very little difference was observed between the groups on test scores, and most of these differences were not in the predicted direction. In fact, those students who received the most training received the lowest scores.

The fact that differences were not found is even more perplexing in light of teachers' very positive response to the program materials. Most teachers who used the materials during this year plan to continue using the materials in the future and felt that the materials had improved their students attitude and increased performance on the standardized achievement test. However, the fact remains that none of these perceived differences were evident on objective measures for which data were collected. These contradictions with previous research and with teachers' perceptions of benefit suggest that further evaluation of the materials developed in this project should be conducted before final conclusions are drawn.

A number of facts learned during this project should assist in making further evaluation as meaningful as possible. Summarized in the remaining two sections are potential explanations for why the training materials and procedures were not as effective as they might have been and factors which should be taken into account in conducting further research and evaluation.

Implementation Factors Possibly Related to Results

Even though previous research suggests that an intervention of the type delivered in this project should have led to substantial differences between groups on students' performance and attitudes during standardized achievement tests; few such differences were noted, and those that were, were relatively small. A closer examination of the data and the implementation procedures suggests a number of possible explanations. None of these can "be proven" as a causal agent in the results that were obtained. They are presented here to provide the reader with a more complete context in interpreting the results that have been presented as well as providing the background for further research and evaluation.

---

[1] As noted in the Procedures Section, there were some classrooms where the training was implemented less well or where some training was not delivered. However, excluding these classrooms from the analyses as reported in the Results Section made no practical difference.

Amount of practice per concept. As noted in the implementation section, over 50 different concepts were presented to students in the filmstrips. Many of these concepts are reasonably complex such as elimination, differentiating between correct answers and look-alikes or sound-alikes, deductive reasoning, and checking work. During the filmstrips, students were provided with a certain amount of practice in each of these concepts. The practice tests which were given on a different day from the filmstrip provided additional opportunity for practicing these concepts; even though practice tests were not designed to give explicit practice with each concept taught in an associated filmstrip.

Because every filmstrip lasted from 20 to 40 minutes and contained four or more major concepts, it is possible that students did not have enough practice time to really master each of the concepts taught in the filmstrips. Breaking the filmstrips into smaller pieces and providing more opportunity for practice may result in more effective instruction. However, this possibility must be considered in light of the fact that the training provided in this intervention was already much more substantial and contained as much or more practice than most other efforts at training students in test-taking skills reported in the literature.

Reinforcement procedures. Based on previous work by Taylor and White (1982), as well as the previous research reported in the review of literature in Chapter II, one component of Experimental Group I was designed to motivate students to try their best on tests through using a self-charting procedure. Taylor and White (1982) demonstrated that paying students to perform better than predicted from their pretest scores on standardized achievement tests resulted in approximately half a standard deviation difference between experimental and control groups. Because it was unacceptable to continue paying students for their performance on standardized achievement tests, the self-charting procedures associated with practice tests were designed to determine if increased motivation during practice tests could be generated and if such motivation would generalize to the standardized achievement test.

Unfortunately, the design of the experiment did not allow for the effects of the reinforcement procedures to be estimated separately from the effects of training students, training teachers, and participating in the practice tests. However, teacher feedback indicated that the reinforcement procedures were the weakest part of the program and were sometimes confusing for students. As noted in the Procedures Section, the self-charting procedures used to motivate students are reasonably complex to implement. Some teachers noted in the debriefing that these particular procedures did not seem to be motivating for students even on the practice tests. If the motivational procedures were ineffective on the practice tests, the probability of increased motivation on the actual standardized achievement test is virtually nonexistent. There is even a possibility that instead of being motivating, the so-called reinforcement procedures were actually a negative influence for children in Experimental Group I.

Format of practice tests. The construction of the practice tests was done so that the tests paralleled as closely as possible the standardized achievement test that the students would be taking in the springtime. The last several practice tests contained several items for each subtest included on the reading portion of that district's standardized achievement test. A single practice test (which took a maximum of 30 minutes) contained up to 7 subtests depending on the district. For each subtest, there was a sample item and directions. Consequently, much of the practice test time was spent giving directions and reviewing sample items. This may have reduced the effectiveness of the practice test because instead of spending most of the time practicing taking tests, students were spending substantial time listening to directions and going over sample items. For the first several practice tests, this was not a problem because the tests were relatively short. By the time the problem was recognized in the longer practice tests, it was too late to redesign the tests so that any given practice test would have only one or two subtests.

Heterogeneity or non-comparability of classrooms within each experimental group. Although classes were randomly assigned to each of the treatment groups from a larger population who had expressed their willingness to participate in the program, there is a possibility that sampling fluctuation could have resulted in non-comparable groups. Because pretest data were not available, it is impossible to check this possibility directly. However, several other sources of data were examined. First, the results of standardized achievement tests for third grade students in the same schools were examined. These data are reported in Table 43. As can be seen, the results are reasonably comparable. Any bias which does exist would have contributed to higher scores for Experimental Group I which was the lowest-scoring group on almost all of the achievement test scores.

It should be noted, however, that the fluctuation in the third grade scores was as great or greater as the differences in scores observed between

Table 43

Third Grade Standardized Achievement Test Scores from 1981-82 Year

GROUP I	TOTAL		MATH		READING		GROUP II	TOTAL		MATH		READING		CONTROL GROUP	TOTAL		MATH		READING	
	GE	%ile	GE	%ile	GE	%ile		GE	%ile	GE	%ile	GE	%ile		GE	%ile	GE	%ile	GE	%ile
Wellsville C	4.7	83	4.8	85	4.6	74	Park C	4.0	59	4.1	68	3.8	52	Millville C	4.0	59	3.9	59	3.9	55
Westside N	4.1	74	4.4	86	3.8	54	Lewiston C	4.0	61	3.9	58	4.1	61	Summit C	4.0	60	4.1	67	3.9	54
Santaquin N	3.9	58	3.5	27	4.1	68	Wilson N	3.9	58	3.7	45	4.0	61	Larsen N	4.7	74	3.7	45	4.2	75
Hillsdale G	4.1	58	4.3	58	4.3	60	Goshen N	3.6	33	3.4	20	3.7	36	Taylor N	4.4	91	4.0	73	4.3	81
Lincoln G	4.2	64	4.7	70	4.4	60	W. Hills G	4.2	66	4.6	68	4.5	62	Brookside N	4.0	66	3.7	45	3.9	53
West Kearns G	4.1	56	4.3	58	4.0	54	Stansbury G	3.9	50	3.9	50	4.0	54	Lake Ridge G	4.1	60	4.2	56	4.4	60
Redwood G	4.4	72	5.0	78	4.6	66	S. Kearns G	4.1	56	4.3	60	4.2	56	Roosevelt G	3.9	50	3.8	48	3.9	50
														Woodrow Wilson G	4.6	76	4.9	77	4.8	70
Mean Scores $\bar{X}$	4.2	66.4	4.4	66.0	4.3	62.3		4.0	54.7	4.0	52.7	4.0	54.6		4.1	67.0	4.0	58.8	4.2	62.3

Note: Letter after school name indicates District: C = Cache, N = Nebo, G = Granite.

BEST COPY AVAILABLE

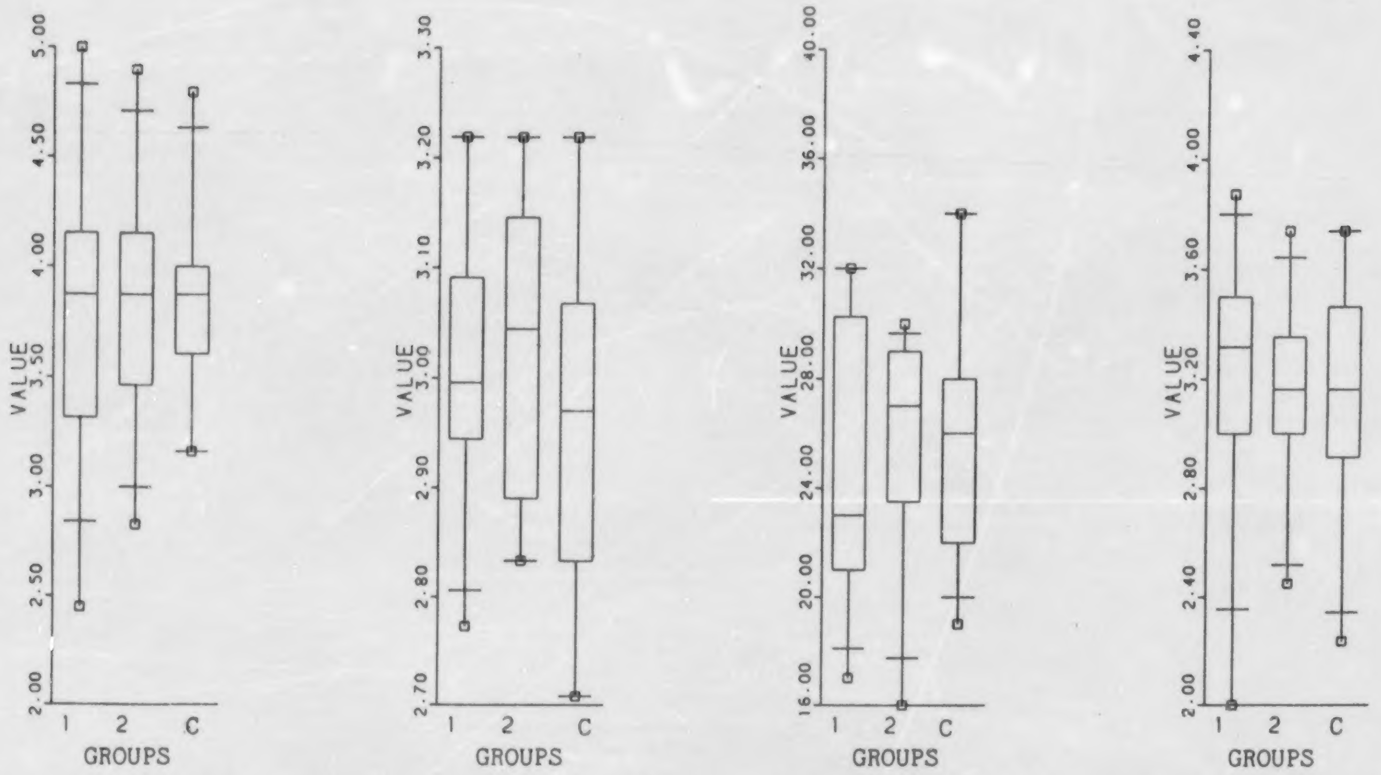
the three second grade experimental groups. Thus, although these data suggest that such fluctuation is easily possible, the direction of the fluctuation at the third grade leads one to believe that such fluctuation is not a primary explanation for the results observed with the second grade students participating in the study.

A related possibility is that a small number of "outlier" teachers in Groups E1 or E2 unduly affected the data. This possibility was examined by constructing Box and Whisker plots for each of the dependent variables using teachers as the unit of analysis. If there were "outliers" in Groups E1 or E2, this would show up by extremely long tails for either Groups E1 or E2 and the hash mark for the 5th percentile being substantially nearer the Box for Control Group than for Groups E1 or E2. These data are presented in Figure 6. As can be seen from the data for the achievement test scores, there are no major differences between the groups. It is apparent that for all of the groups on several dependent measures such as quality of test administration, and time on task for teachers, the distributions are negatively skewed. However, these data do not suggest that a small number of teachers are unduly affecting the scores of Groups E1 or E2 which might lead to a misinterpretation of the data.

"John Henry" effect. The "John Henry effect" suggests that control group teachers who know they are being compared to an experimental treatment will try harder and thus perform better than they would under typical conditions. If such extra efforts were present on the part of control group teachers, the results of the experiment would be invalidated. Because all of the control group teachers were aware of the general nature of the study being conducted, it is possible that such a "John Henry effect" existed. However, the results

Figure 6

Box and Whisker Diagrams for Dependent Variables
Using Teacher as Unit of Analysis



Teacher Attitude
(Opinion About Tests)
[Square root transformation]

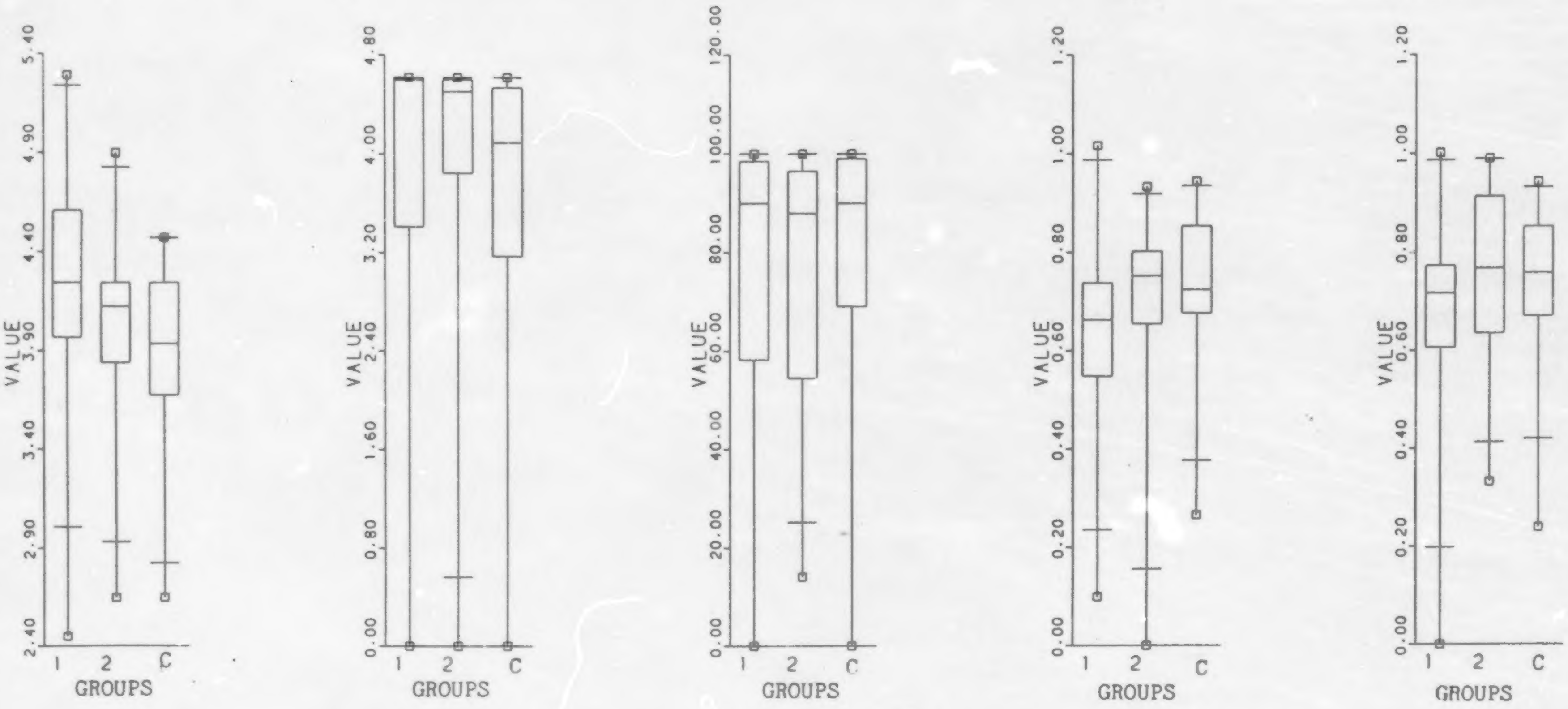
Teacher Attitude
(Feeling About Administration)
[Log transformation]

Teacher Attitude
(Usefulness of Tests)
[Raw scores]

Teacher Attitude
(Should Use Be Increased?)
[Square root transformation]

Figure 6 (cont'd)

Box and Whisker Diagrams for Dependent Variables
Using Teacher as Unit of Analysis



Teacher Attitude
(Students' Feelings)
[Square root transformation]

Teacher On-Task
(Teacher-Directed)
[Log transformation]

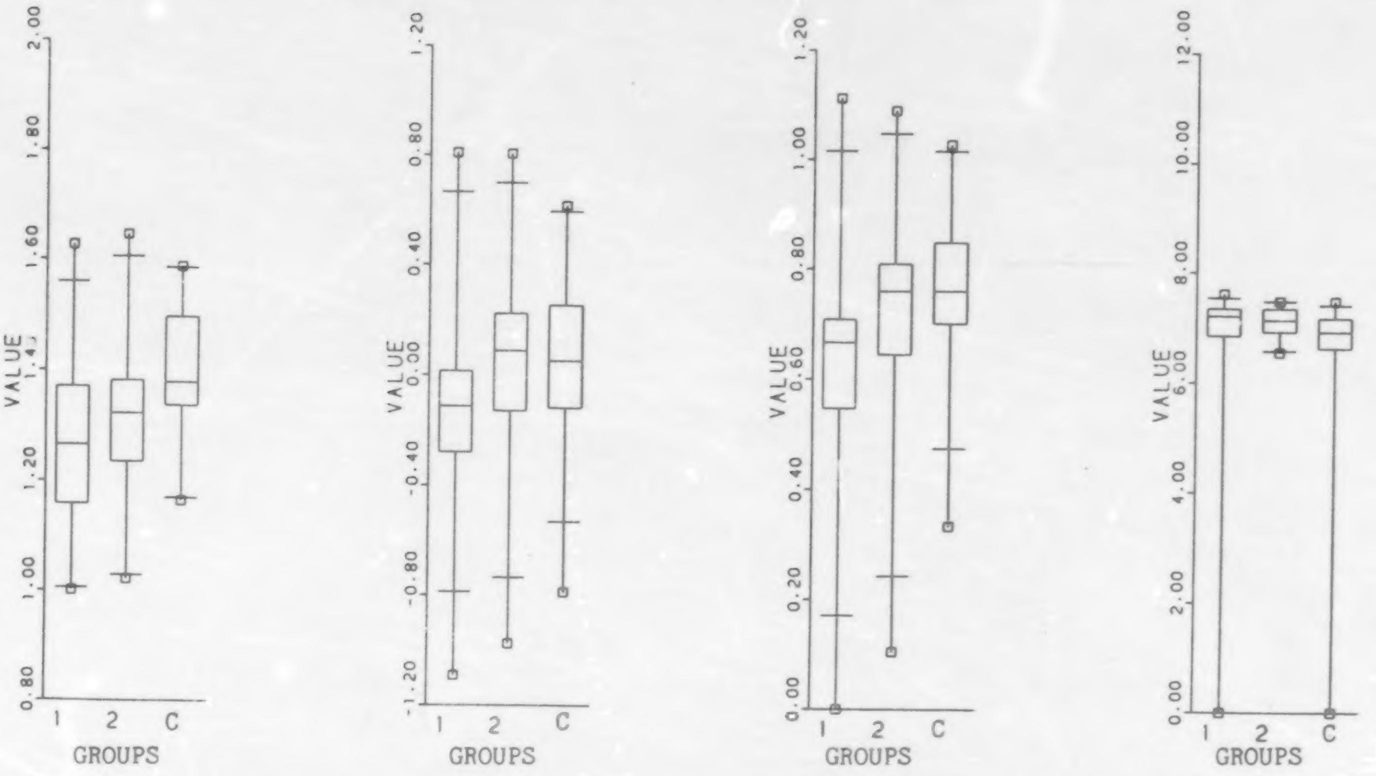
Teacher On-Task
(Student-Directed)
[Raw scores]

Student On-Task
(Teacher-Directed)
[Log transformation]

Student On-Task
(Student-Directed)
[Log transformation]

Figure 6 (cont'd)

Box and Whisker Diagrams for Dependent Variables
Using Teacher as Unit of Analysis



Achievement Test (Math)
[Square root transformation]

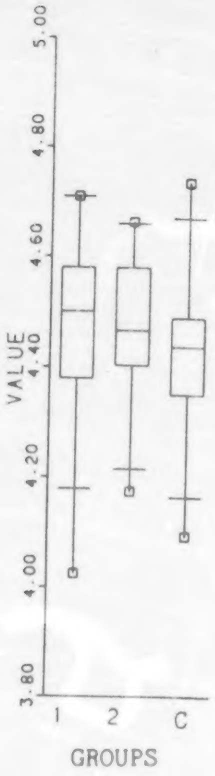
Achievement Test (Reading)
[Raw scores]

Achievement Test
(Total)
[Log transformation]

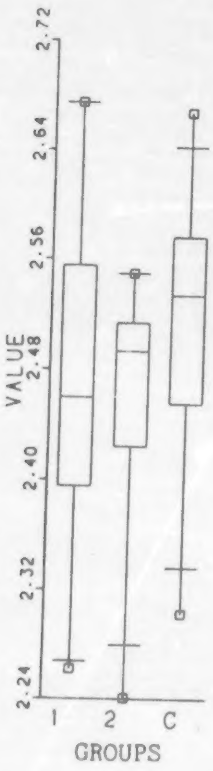
Quality of Test
Administration
[Square root transformation]

Figure 6 (cont'd)

Box and Whisker Diagrams for Dependent Variables Using Teacher as Unit of Analysis



Teacher Attitude (Total Scale)
[Log transformation]



Student Attitude Towards Testing
[Log transformation]

BEST COPY AVAILABLE

of the debriefing interviews and the contact with control group teachers in preparation for collecting the observational data suggests that such an effect is unlikely.

Timing of implementation. The original schedule for implementation was that teachers would show a filmstrip approximately every two weeks with practice tests in between. Unfortunately, production schedules for the filmstrips had to be altered due to unforeseen circumstances, and the first three filmstrips were spread out over three months with the last five filmstrips occurring in a period of only eight weeks. In addition, as noted in the Procedures Section, none of the teachers in Nebo District showed Filmstrip #7. It may have been that such irregularities in the implementation attenuated gains that might have resulted from the filmstrips and practice tests. Teachers in the debriefing interviews did not feel this was a serious problem, but it is hard to estimate what effect the scheduling irregularities may have had on children.

Does better test administration lead to higher scores? Previous research reported by Taylor and White (1982) demonstrated that students who took standardized tests from teachers who had been trained in proper test administration obtained higher test scores than students who had not. In some ways, it is logical that better test administration would result in higher scores. For example, high quality test administration would mean that teachers would give better directions, would be better at keeping students on task, and would prepare students better for taking the test. All of these things would probably lead to higher test scores. Alternatively, however, better test administration could lead to lower scores if the better test administration reduced cheating and eliminated unfair teacher assistance or hints. If, as a result of training in test-taking skills and better test administration procedures, students' scores improve more than would have been predicted from a "practice effect", one can be relatively confident that the previous scores were not valid indicators of what the student knows. However, if students receive lower test scores, it is unclear whether the latter test scores are more or less valid. Determining the degree to which scores are valid is a time-consuming and complex process which, given these results, lies beyond the scope of the project.

Student fatigue or overconfidence. It is possible that the student training implemented in E1 and E2 actually resulted in students becoming fatigued with taking tests or so overconfident in their ability to take tests that they did not perform as well on the actual standardized achievement test. Particularly because of the scheduling problems which resulted in the students receiving 4-5 reasonably long practice tests in the last two months, students in E1 and E2 may have become "desensitized" to the importance of the standardized achievement test and performed below their true level of achievement.

Suggestions for Future Research

The results from this study do not demonstrate that the use of these materials results in more valid standardized achievement test scores or better performance or attitudes on the part of students. According to the measures designed for this study, they do suggest that teachers who have participated in the project have better attitudes toward standardized tests and are more capable test administrators. However, because the results concerning student performance contradict the conclusions of previous research in related areas, and because of teachers' perceptions that the materials were beneficial, it is suggested that the results of this project not be taken as the final word. Further research should be conducted. In conducting that research, several suggestions are made based on the data collected during this study.

1. Practice tests should be revised so that less time is spent giving directions and going over sample items. This could be accomplished by having each practice test include only one or two subtests. In addition, it might be worthwhile to create a one-to-one correspondence between the concepts taught in the filmstrip and practiced in the practice test.
2. Future studies should be designed so that the effects of reinforcement, teacher training, student training, and practice tests can be examined independently from each other. This type of design would make any results easier to understand. Of course, this type of design requires more students and costs more money (all other things being equal).
3. The filmstrips and student training packages should be redesigned into smaller components, there should be no more than 15 to 20 minutes per training session, and should include additional practice on each of the concepts taught. Also, substantial time should be invested with smaller groups of students going through the filmstrips before testing them on a large population. Such a study should run over several years. One of the main problems with the current project was trying to do extensive curriculum development work while simultaneously conducting a large-scale field study.
4. The "reinforcement" procedures used in the study need to be completely re-examined and perhaps reconceptualized. The impetus for this work was the work reported by Taylor and White (1982) in which students were paid money for their performance above that which would have been predicted from their pretest score on a standardized achievement test. That work should probably be replicated to first

determine whether motivation is as large a factor as it appeared in the Taylor and White study. If indeed it can be demonstrated that motivational factors are a primary consideration in the validity of standardized test scores, then alternative ways of motivating students to perform on standardized tests should be found. The way in which the present study was designed, it was impossible to separately estimate the effects of reinforcement to determine whether or not the procedure was actually reinforcing.

5. The training materials should be tried with children at different grade levels. Second grade children were chosen for this study because we wanted to train students as near as possible to the beginning of their standardized testing experience before they had learned "bad habits" which would have to be unlearned. However, the fact remains that the concepts taught may have been too complex for second graders or that the emphasis on test taking at such an early age may have made them more anxious.

The most important conclusion from these suggestions, however, is that further research is necessary to understand to what degree typically administered standardized achievement tests are valid and useful for the purposes for which they are usually used. The materials developed in this project represent an important beginning. As they are used further and more data are collected, we will be able to better understand the degree to which results from standardized tests should and can be used to make programming, evaluation, and placement decisions for primary grade children.

REFERENCES

Aiken, L. R., & Williams, E. N. Effects of instructions, option keying, and knowledge of test material on seven methods of scoring two-option items. *Educational and Psychological Measurement*, 1978, *38*, 53-59.

Alker, H. A., Carlsen, J. A., & Hermann, M. G. Multiple choice questions and student characteristics. *Journal of Educational Psychology*, 1969, *60*, 231-243.

Allen, G. J. The behavioral treatment of test anxiety: Recent research and future trends. *Behavior Therapy*, 1972, *3*, 253-262.

Alpert, R., & Haber, R. N. Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology*, 1960, *61*, 207-215.

Anastasi, A. *Psychological testing* (4th ed.). New York: Macmillan, 1976.

Arnold, S. T. The effects of two types of group test anxiety management techniques on college students' underachievement (Doctoral dissertation, University of Indiana, 1979), *Dissertation Abstracts International*, 1979, *40*, 668A. (University Microfilms No. 7916882)

Axelrod, S. *Behavior modification for the classroom teacher*. New York: McGraw-Hill, 1972.

Axelrod, S., Hall, R. V., & Tams, A. *A comparison of two common seating arrangements in classroom settings*. Paper presented at annual meeting of Kansas Symposium on Behavior Analysis in Education, Lawrence, Kansas, May 1972.

Ayllon, T., & Kelly, K. Effects of reinforcement on standardized test performance. *Journal of Applied Behavior Analysis*, 1972, *5*, 477-484.

Back, R. D. The effects of pretest exposure on sex of examiner influence on the Wechsler Intelligence Scale for Children (Doctoral dissertation, University of Arkansas, 1978). *Dissertation Abstracts International*, 1979, *40*, 1348-B. (University Microfilms No. 7919223)

Baer, R. D. *Effects of reinforcement on intelligence test behavior as a function of test administered and sex of subject*. Unpublished doctoral dissertation, Utah State University, 1978.

Baird, L. L. *What graduate and professional school students think about admissions tests*. East Lansing, Michigan: Michigan State University, 1977. (ERIC Document Reproduction Service No. 157903)

Bath, J. A. Answer-changing behavior on objective examinations. *Journal of Educational Research*, 1967, *61*, 105-107.

Bell, F. O., Hoff, A. L., & Hoyt, K. B. Answer sheets do make a difference. *Personnel Psychology*, 1964, *17*, 65-71.

Benson, G. G. E. The effect of immediate item feedback on the reliability and validity of verbal ability test scores (Doctoral dissertation, Florida State University, 1979). *Dissertation Abstracts International*, 1980, *40*, 5410A. (University Microfilms No. 8008637)

Berrien, F. K. Are first impressions best on objective tests? *School and Society*, 1939, *50*, 319-320.

Bridgeman, B. Effects of test score feedback on immediately subsequent test performance. *Journal of Educational Psychology*, 1974, *66*, 62-66.

Butler, P. T. The effects of item-difficulty sequences and test-taking anxiety-reaction types on mathematics performance (Doctoral dissertation, University of Virginia, 1979). *Dissertation Abstracts International*, 1980, *40*, 5013A-5014A. (University Microfilms No. 8004659)

Callenbach, C. The effects of instruction and practice in content-independent test-taking techniques upon standardized reading test scores of selected second-grade students. *Journal of Educational Measurement*, 1973, *10*, 25-29.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand-McNally, 1963.

Cashen, V. M., & Ramseyer, G. C. The use of separate answer sheets by primary aged children. *Journal of Educational Measurement*, 1969, *6*, 155-158.

Cassel, R. N. Basic assumptions underlying inferences for the usual psychological test scores. *Journal of Employment Counseling*, 1969, *6*, 56-59.

Cieutat, V. J., & Flick, G. L. Examiner differences among Stanford-Binet items. *Psychological Reports*, 1967, *21*, 613-622.

Cohen, J. *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press, 1977.

Coleman, J. I. The effect of examiner warmth on test anxiety and performance during group intelligence testing of latency aged females (Doctoral dissertation, Boston University School of Education, 1978). *Dissertation Abstracts International*, 1978, *38*, 7148A-7149A. (University Microfilms No. 7808056)

Competency tutoring program. New York: Public Education Association, 1979. (ERIC Document Reproduction Service No. ED 175 951)

Couch, J. V., Turner, W. E., & Garber, T. B. *Self statements, test anxiety and academic achievement: A correlational analysis*. Paper presented at the meeting of the Association for Behavioral Analysis, Dearborn, 1979.

Crehan, K. D., Gross, L. J., Koehler, R. A., & Slakter, M. J. *Developmental aspects of test-wiseness*. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Crehan, K. D., Gross, L. J., Koehler, R. A., & Slakter, M. J. Developmental aspects of test-wiseness. Educational Research Quarterly, 1978, 3, 40-44.

Crehan, K. D., Koehler, R. A., & Slakter, M. J. Longitudinal studies of test-wiseness. Journal of Educational Measurement, 1974, 11, 209-212.

Davis, W. E., Peacock, W., Fitzpatrick, P., & Mulhern, M. Examiner differences, prior failure, and subjects' arithmetic scores. Journal of Clinical Psychology, 1969, 25, 178-180.

Diamond, J. J., & Evans, W. J. An investigation of the cognitive correlates of test-wiseness. Journal of Educational Measurement, 1972, 9, 145-150.

Diamond, J. J., Ayrer, J., Fishman, R., & Green, P. Are inner city children test-wise? Journal of Educational Measurement, 1976, 14, 39-45.

Doffenbacher, J. L. Worry, emotionality, and task-generated interference in test anxiety: An empirical test of attentional theory. Journal of Educational Psychology, 1978, 70, 248-254.

Downey, G. W. Is it time we started teaching children how to take tests? The American School Board Journal, 1977, 164, 27-31.

Eaton, W. O. Profile approach to longitudinal data: Test anxiety and success-failure experience. Developmental Psychology, 1979, 15, 344-345.

Ebel, R. L. The paradox of educational testing. Paper presented at the meeting of the National Council on Measurement in Education, East Lansing, 1976.

Ebel, R. L., & Damrin, D. E. Tests and examinations. In C. Harris (Ed.), Encyclopedia of Educational Research. New York: Macmillan Co., 1960.

Erickson, M. E. Test sophistication: An important consideration. Journal of Reading, 1972, 16, 140-146.

Exner, J. E. Variations in WISC performances as influenced by differences in pretest rapport. Journal of Genetic Psychology, 1966, 74, 299-306.

Fenton, N. An objective study of student honesty during examinations. School and Society, 1927, 26, 341-344.

Ferrell, G. The relationship of scores on a measure of test-wiseness to performance on teacher-made objective achievement examinations and on standardized ability and achievement tests, to grade point average, and to sex for each of five high school samples (Doctoral dissertation, University of Southern California, 1972). Dissertation Abstracts International, 1972, 33, 1510A. (University Microfilms No. 72-26, 013)

Ferrell, G. Development and use of a test of test-wiseness. Paper presented at the annual meeting of the Western College Reading Association, Denver, March 1977.

Friedman, M. H. Test anxiety: An investigation of its nature and remediation (Doctoral dissertation, University of Texas at Austin, 1979). *Dissertation Abstracts International*, 1979, *40*, 1363B-1364B. (University Microfilms No. 7920119)

Fueyo, V. Training test-taking skills: A critical analysis. *Psychology in the Schools*, 1977, *14*, 180-184.

Gaffney, R. F., & Maguire, T. O. Use of optically scored test answer sheets with young children. *Journal of Educational Measurement*, 1971, *8*, 103-106.

Gibb, B. G. *Test-wiseness as secondary cue responses*. Unpublished doctoral dissertation, Stanford University, 1964.

Glass, G. V. Integrating findings: The meta-analysis of research. *Review of Research in Education*, 1977, *5*, 351-379.

Glass, G. V., & Smith, M. L. Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis*, 1979, *1*, 2-16.

Hammerton, M. The guessing correction in vocabulary tests. *British Journal of Educational Psychology*, 1965, *35*, 249-251.

Heim, A. W., & Wallace, J. G. The effects of repeatedly retesting the same group on the same intelligence test: I, Normal Adults. *Quarterly Journal of Experimental Psychology*, 1949, *1*, 151-159.

Hill, K. T. Social reinforcement as a function of test anxiety and success-failure experiences. *Child Development*, 1967, *38*, 723-737.

Hill, K. T., & Eaton, W. O. The interaction of test anxiety and success-failure experiences in determining children's arithmetic performance. *Developing Psychology*, 1977, *13*, 205-211.

Hill, K. T., & Sarason, S. B. The relation of test anxiety and defensiveness to test and school performance over the elementary school years. A further longitudinal study. *Monographs of the Society for Research in Child Development*, Serial No. 104, 1966, *31* (Whole No. 2).

Holmes, E. Reading guided by questions versus careful reading and re-reading without questions. *School Review*, 1931, *39*, 361-371.

Holroyd, K. A. Cognition and desensitization in the group treatment of test anxiety. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 991-1001.

Hoyt, C. Test reliability obtained by analysis of variance. *Psychometrika*, 1941, *6*, 153-160.

Huck, S. W. Test performance under the condition of known item difficulty. *Journal of Educational Measurement*, 1978, *15*, 53-58.

Jensen, A. *Bias in mental testing*. New York: Free Press, 1980.

The Joint Dissemination Review Panel. *Ideabook*. Washington, D.C.: Superintendent of Documents, U.S. Government Printing Office, October 1977.

Katz, I., Henchy, T., & Allen, H. Effects of race of tester, approval-disapproval, and need on negro children's learning. *Journal of Personality and Social Psychology*, 1968, *8*, 38-42.

Katz, I., Roberts, S. O., & Robinson, J. M. Effects of task difficulty, race of administrator, and instructions on digit-symbol performance of negroes. *Journal of Personality and Social Psychology*, 1965, *2*, 53-59.

Kelley, T. L. Cumulative significance of a number of independent experiments: Reply to D. E. Traxler and R. N. Hilkert. *School and Society*, 1943, *57*, 482-484.

Kestenbaum, J. M., & Weiner, B. Achievement performance related to achievement motivation and test anxiety. *Journal of Consulting and Clinical Psychology*, 1970, *34*, 343-344.

Kirkland, M. C. The effects of tests on students and schools. *Review of Educational Research*, 1971, *41*, 303-349.

Krueck, T. Personal communication, March 17, 1981.

Kubiszyn, T. W. The effects of knowledge of item difficulty, IQ and test anxiety on classroom test performance in undergraduate females (Doctoral dissertation, University of Texas at Austin, 1979). *Dissertation Abstracts International*, 1979, *40*, 1362A. (University Microfilms No. 7920148)

Lazarus, R. S., & Eriksen, C. W. Effects of failure stress upon skilled performance. *Journal of Experimental Psychology*, 1972, *43*, 100-105.

Lehman, H. C. Does it pay to change initial decisions in a true-false test? *School and Society*, 1928, *28*, 456-458.

Lent, R. W., & Russell, R. K. Treatment of test anxiety by cue-controlled desensitization and study-skills training. *Journal of Consulting Psychology*, 1978, *25*, 217-224.

Light, R. J., & Smith, D. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, 1971, *41*, 429-471.

Loret, P. G. Personal communication, August 29, 1980.

Lowe, M. L., & Crawford, C. C. First impression versus second thought in true-false tests. *Journal of Educational Psychology*, 1929, *20*, 192-195.

Lynch, D. L., & Smith, B. C. *To change or not to change item responders when taking tests: Empirical evidence for test takers*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1972.

Magnusson, D. *Test theory*. Reading, Mass.: Addison Wesley, 1967.

Mandler, G., & Sarason, S. B. A study of anxiety and learning. *Journal of Abnormal and Social Psychology*, 1952, *47*, 166-173.

Marlett, N. J., & Watson, D. Test anxiety and immediate or delayed feedback in a test-like avoidance task. *Journal of Personality and Social Psychology*, 1968, *8*, 200-203.

Masling, J. The influence of situational and interpersonal variables in projective testing. *Psychological Bulletin*, 1960, *57*, 65-85.

Mathews, C. D. Erroneous first impressions on objective tests. *Journal of Educational Psychology*, 1929, *20*, 280-286.

McCandless, B. R., & Castaneda, A. Anxiety in children, school achievement and intelligence. *Child Development*, 1956, *27*, 379-381.

McCarthy, D. A study of the reliability of the Goodenough drawing test of intelligence. *Journal of Psychology*, 1944, *18*, 201-216.

McCoy, N. Effects of test anxiety on children's performance as a function of instructions and type of task. *Journal of Personality and Social Psychology*, 1965, *2*, 634-641.

McGaw, B., & White, K. *Meta-analysis of empirical research*. AERA Research Training Presession at the meeting of the American Educational Research Association, Los Angeles, April 1981.

McKenna, B. *Finding and recommendations, interim report, NEA task force on testing*. Washington, D.C.: National Education Association, May 29, 1973.

Meichenbaum, D. H. Cognitive modification of test anxious college students. *Journal of Consulting and Clinical Psychology*, 1972, *39*, 370-380.

Mercer, M. Answer changing and students' test scores (Doctoral dissertation, Rutgers University, The State University of New Jersey, New Brunswick, 1978). *Dissertation Abstracts International*, 1979, *40*, 214A-215A. (University Microfilms No. 7914126)

Messick, S., & Anderson, S. Educational testing, individual development, and social responsibility. *Counseling Psychologist*, 1970, *2*, 80-88.

Millikin, J. L. Some correlates of test-wiseness among high school students (Doctoral dissertation, Texas A & M University, 1975). *Dissertation Abstracts International*, 1976, *36*, 5155A. (University Microfilms No. 76-3653)

Millman, J., Bishop, C. H., & Ebel, R. An analysis of test-wiseness. *Educational and Psychological Measurement*, 1965, *25*, 707-726.

Millman, J., & Pauk, W. *How to take tests*. New York: McGraw-Hill, 1967.

Millman, J., & Setijadi, I. A comparison of the performance of American and Indonesian students on three types of test items. *Journal of Educational Measurement*, 1966, *59*, 273-275.

Mini-test. New York: Educational Solutions, Inc., 1979.

Moore, J. C., Schutz, R. E., & Baker, R. L. The application of a self-instructional technique to develop a test-taking strategy. *American Educational Research Journal*, 1966, *3*, 13-17.

Newland, T. E. Assumptions underlying psychological testing. *Journal of School Psychology*, 1973, *11*, 316-322.

Nunn, G. D. *Test anxiety and perception of self: A correctional study*. Manhattan, Kansas: Kansas State University, 1976. (ERIC Document Reproduction Service No. ED 169 426)

Oakland, T. The effects of test-wiseness materials on standardized test performance of preschool disadvantaged children. *Journal of School Psychology*, 1972, *10*, 355-360.

Orfanos, J. D. Effect of "game" versus "test" portrayal on performance of a complex cognitive task. *Measurement and Evaluation in Guidance*, 1979, *12*, 121-124.

Osler, S. F. Intellectual performance as a function of two types of psychological stress. *Journal of Experimental Psychology*, 1954, *47*, 115-121.

Paquin, M. J. The effects of pupil self graphing on academic performance. *Education and Treatment of Children*, 1978, *1*, 5-16.

Parker, K. E. The differential effectiveness of specific treatments on the worry-emotionally components of test anxiety (Doctoral dissertation, George Peabody College for Teachers, 1979). *Dissertation Abstracts International*, 1980, *41*, 363-B. (University Microfilms No. 8016114)

Paul, G. L., & Eriksen, C. W. Effects of test anxiety on "real-life" examinations. *Journal of Personality*, 1964, *32*, 480-494.

Pease, G. R. Sex differences in algebraic ability. *Journal of Educational Psychology*, 1930, *21*, 712-714.

Piersel, W. C., Brody, G. H., & Kratochwill, T. R. A further examination of motivational influences on disadvantaged minority group children's intelligence test performance. *Child Development*, 1977, 48, 1142-1145.

Rechebei, E. Personal communication, September 2, 1980.

Reichenberg-Hackett, D. Changes in Goodenough drawings after a gratifying experience. *American Journal of Orthopsychiatry*, 1953, *23*, 501-517.

Reile, P. J., & Briggs, L. J. Should students change their initial answers on objective-type tests?: More evidence regarding an old problem. *Journal of Educational Psychology*, 1952, *43*, 110-115.

Roberts, O. *Practice effect or test wiseness*. Unpublished manuscript, 1979. (Available from RMC Research Corporation, Mountain View, California)

Rodger, A. G. The application of six group intelligence tests to the same children, and the effects of practice. British Journal of Educational Psychology, 1936, 6, 291-305.

Rowley, G. L. Which examinees are most favoured by the use of multiple choice tests? Journal of Educational Measurement, 1974, 11, 15-23.

Ruebush, B. K. Children's behavior as a function of anxiety and defensiveness. Unpublished doctoral dissertation, Yale University, 1960.

Sacks, E. L. Intelligence scores as a function of experimentally established social relationships between child and examiner. Journal of Abnormal Social Psychology, 1952, 47, 354-358.

Sarason, I. G. Test anxiety, general anxiety, and intellectual performance. Journal of Consulting Psychology, 1957, 21, 485-490.

Sarason, I. G. Individual differences, situational variables, and personality research. Journal of Abnormal and Social Psychology, 1962, 65, 376-380.

Sarason, I. G. Test anxiety and intellectual performance. Journal of Abnormal and Social Psychology, 1963, 66, 73-75.

Sarason, I. G. Test anxiety and cognitive modeling. Journal of Personality and Social Psychology, 1973, 28, 58-61.

Sarason, I. G. The test anxiety scale: Concept and research. In C. D. Spielberger & I. G. Sarason (Eds.), Stress and anxiety (Vol. 5). New York: Hemisphere/Wiley, 1978.

Sarason, I. G., & Minard, J. Interrelationships among subject, experimenter, and situational variables. Journal of Abnormal and Social Psychology, 1963, 67, 87-91.

Sarason, I. G., & Palola, E. G. The relationship of test and general anxiety, difficulty of task, and experimental instructions to performance. Journal of Experimental Psychology, 1960, 59, 185-191.

Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., & Ruebush, B. K. Anxiety in elementary school children. New York: Wiley, 1960.

Sarnacki, R. E. An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 1979, 49, 252-279.

Sattler, J. M., & Theve, F. Procedural, situational, and interpersonal variables in individual intelligence testing. Psychological Bulletin, 1967, 68, 347-360.

Scheib, J. E. Convergent and discriminate validation of test-wiseness and risk-taking on objective examinations (Doctoral dissertation, University of Pennsylvania, 1979). Dissertation Abstracts International, 1979, 40, 1422A-1423A. (University Microfilms No. 7919509)

Seitz, V., Abelson, W. D., Levine, E., & Zigler, E. Effects of place of testing on the Peabody Picture Vocabulary Test scores of disadvantaged Head Start and non-Head Start children. *Child Development*, 1975, *46*, 481-486.

Shannon, A. J. *Some effects of methods of standardized reading achievement test administration and score interpretation on senior high students' attitude toward reading*. Unpublished doctoral dissertation, Marquette University, 1978.

Sherriffs, A. C., & Boomer, D. S. Who's penalized by the penalty for guessing? *Journal of Educational Psychology*, 1954, *45*, 81-90.

Slakter, M. J. The penalty for not guessing. *Journal of Educational Measurement*, 1968, *5*, 141-145.

Slakter, M. J., Koehler, R. A., & Hampton S. H. Learning test-wiseness by programmed texts. *Journal of Educational Measurement*, 1970, *7*, 247-254.

Steininger, M., Johnson, R. E., & Kirts, D. K. Cheating on college examinations as a function of situationally aroused anxiety and hostility. *Journal of Educational Psychology*, 1964, *55*, 317-324.

Stoneman, Z., & Gibson, S. Situational influences on assessment performance. *Exceptional Children*, 1978, *45*, 166-169.

Strang, H. R., Bridgeman, B., & Carrico, M. F. Effects of "game" versus "test" task definition for third grade children on three subtests of the Wechsler Intelligence Scale for Children. *Journal of Educational Measurement*, 1974, *11*, 125-128.

Tallmadge, G. K., & Wood, C. T. *User's guide: Title I evaluation and reporting system* (Rev.). Prepared for Department of Education. Mountain View, Ca.: RMC Research Corporation, 1981.

Taylor, C., & White, K. R. The effects of reinforcement and training on group standardized test behavior. *Journal of Educational Measurement*, 1982, *19*, 199-210.

Taylor, P. H. A study of the effects of instructions in a multiple-choice mathematics test. *British Journal of Educational Psychology*, 1966, *36*, 1-6.

Test taking skills kit. Herndon, Virginia: Evaluation and Assessment Service, Inc., 1980.

Thomas, A., Hertzig, M. E., Dryman, I., & Fernandez, P. Examiner effect in IQ testing of Puerto Rican working-class children. *American Journal of Orthopsychiatry*, 1971. *41*, 809-821.

Thorndike, R. L. *Personnel selection: Test and measurement techniques*. New York: John Wiley & Sons, Inc., 1949.

Thorndike, R. L. Reliability. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1951.

Traxler, A. E. Administering and scoring the objective test. In E. F. Lindquist (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1963.

Traxler, A. E., & Hiekert, R. N. Effects of type of desk on results of machine-scored tests. School and Society, 1942, 56, 277-296.

Tryon, G. S. The measurement and treatment of test anxiety. Review of Educational Research, 1980, 50, 343-372.

Tukey, J. W. Exploratory data analysis. Reading, Mass.: Addison/Wesley, 1977.

Tyler, F. T., & Chalmers,T. M. Effect on scores of warning junior high school pupils of coming tests. Journal of Educational Research, 1943, 37, 290-296.

Ullman, L. P., & Krasner, L. (Eds.). Case studies in behavior modification. New York: Holt, Rinehart, & Winston, 1965.

Van Houten, R., & Parsons, S. An analysis of a performance feedback system: The effects of timing and feedback, public posting, and praise upon academic performance and peer interaction. Journal of Applied Analysis of Behavior, 1975, 8, 449-457.

Vernon, P. E. Practice and coaching effects in intelligence tests. Educational Forum, 1954, 13, 269-280.

Vernon, P. E. The determinants of reading comprehension. Educational and Psychological Measurement, 1962, 22, 269-286.

Washburne, J. N. The use of questions in social science material. Journal of Educational Psychology, 1929, 20, 321-359.

Weideman, C. C., & Newens, L. F. The effect of directions preceding true-false and indeterminate-statement examinations upon distributions of test scores. Journal of Educational Psychology, 1933, 24, 97-106.

Weiss, R. H. Effects of reinforcement on the IQ scores of preschool children as a function of initial IQ. Unpublished doctoral dissertation, Utah State University, 1980.

White, K. R., Taylor, C., Eldred, N., & Carcelli, L. State refinements to the ESEA Title I evaluation and reporting system: Utah 1979-1980 project. Logan, Utah: Utah State University, Exceptional Child Center, May 1981.

Wickes, T. A. Examiner influence in a testing situation. Journal of Consulting Psychology, 1956, 20, 23-26.

Willis, J. Effects of systematic feedback and self charting on remedial tutorial programs in reading. Journal of Experimental Education, 1974, 42, 83-85.

Wine, J. Test anxiety and direction of attention. Psychological Bulletin, 1971, 76, 92-104.

Witmer, J. M., Dunham, R. M., & Bornstein, A. V. The effects of verbal approval and disapproval upon the performance of third and fourth grade children on four subtests of the WISC. Journal of School Psychology, 1971, 9, 347-356.

Woodley, K. K. Test-wiseness: A cognitive function? Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Zigler, E., Abelson, W. D., & Seitz, V. Motivational factors in the performance of economically disadvantaged children on the Peabody Picture Vocabulary Test. Child Development, 1973, 44, 294-303.

Appendix A

Materials Related to Review Literature

1. Coding Sheets Used for Reinforcement and Student Training Meta-Analyses
2. Summary Listing of ES by Study for Reinforcement and Student Training Meta-Analyses

Factors Affecting
Standardized Test Results

Reinforcement

Author(s) Abbreviated Title

Code	Item	Description
	1. SUBJECTS	
____	1a. Number of Subjects	1 = 12 - 29 2 = 30 - 100 3 = over 100
____	1b. Mean Age	1 = 4 - 6 2 = 7 - 10 3 = 11 - 23
____	1c. Mean IQ	1 = 43 - 85 2 = 86 - 100 3 = over 100
	2. INDEPENDENT VARIABLE	
____	2a. Reinforcer	1 = money 2 = candy 3 = praise 4 = reproof 5 = token 6 = choice 7 = prize
____	2b. Schedule	1 = immediate-item 2 = immediate-subtest 3 = delayed
____	2c. Contingency	1 = contingent 2 = noncontingent
	3. DEPENDENT VARIABLE	
____	3a. Type of Test	1 = academic 2 = intelligence
____	3b. Administration Unit	1 = individual 2 = group
	4. DESIGN	
____	4a. Type of Design	1 = true experimental 2 = quasi-experimental 3 = pre/post
____	4b. Quality of Design	1 = high 2 = low
____	5. EFFECT SIZE	
____	6. CONCLUSIONS	1 = treatment worked 2 = some question 3 = treatment did not work

Factors Affecting
Standardized Test Results
Student Training

________________________________ ________________________________ ____________

Author(s) Abbreviated Title

Code	Item	Description
	1. SUBJECTS	
____	1a. Number of Subjects	1 = 9 - 49 2 = 50 - 99 3 = 100 - 199 4 = 200 - 705 5 = over 1000
____	1b. Mean Age	1 = 5 - 10 2 = 11 - 14 3 = 15 - 18 4 = 19 - 24 5 = 25 - 40
____	1c. Mean IQ	1 = 65 = 89 2 = 90 - 114 3 = 115 - 120
	2. INDEPENDENT VARIABLE	
____	Type of Training	1 = practice 2 = testwiseness
	3. DEPENDENT VARIABLE	
____	3a. Type of Test	1 = achievement 2 = IQ
____	3b. Administration Unit	1 = individual 2 = group
	4. DESIGN	
____	4a. Type of Design	1 = true experimental 2 = quasi-experimental 3 = pre/post
____	4b. Quality of Design	1 = high 2 = low
____	5. EFFECT SIZE	
____	6. CONCLUSIONS	1 = treatment worked 2 = some question 3 = treatment did not work

SUMMARY OF DATA FROM STUDIES ON REINFORCING TESTING BEHAVIOR

ID	ES	Quality	IQ
01	72	2	47
	66	2	93
	92	2	43
02	35	2	106
	- 25	2	106
03	11	1	100
04	- 20	1	119
	- 26	1	102
	269	1	80
	- 03	1	119
	15	1	96
	23	1	78
05	69	1	100
	14	1	100
06	165	1	82
07	25	2	99
08	08	2	100
09	87	2	102
10	95	1	63
11	81	1	65
	79	1	65
12	45	1	90
	23	1	90
	54	1	90
	12	1	90
	16	1	90
	41	1	90
13	11	1	100
	79	1	100
	16	1	100
14	23	2	115
15	12	1	100
	38	1	100
	160	1	100
16	06	1	108
	06	1	108
	06	1	108
17	29	2	114
	95	2	94
	112	2	76
18	136	2	108

SUMMARY OF DATA FROM STUDIES ON TRAINING STUDENTS IN TW

ID	ES	Quality
01	14	1
	16	1
02	78	2
03	69	1
04	39	2
05	404	2
06	197	2
07	67	2
	49	2
08	84	1
09	233	2
	- 27	2
10	20	1
11	53	2
12	32	2
	02	2
13	08	2
14	13	2
	- 08	2
15	05	1
	05	1
16	72	1
17	54	2
	18	2
	12	2
18	74	2
	97	2
19	31	2
	43	2
20	78	2
21	28	1
	13	1
22	13	2
	37	2
	29	2
23	36	1
	15	1
24	30	2
25	03	1
26	06	1
27	20	1
28	119	2
	58	2
	82	2
	72	2
29	109	1
30	23	1
31	183	2
32	84	2
	73	2
	69	2
33	56	2
34	48	2
35	142	2
	78	2
	138	2
	126	2
	138	2
36	110	2
37	21	2
	24	2
	21	2

Appendix B

Materials Related to Development of Filmstrips

1. Letter from Northwest Regional Educational Laboratory about Frequency with which Different Tests Are Used by Title I Projects in Utah
2. Information on Frequency with which Different Tests Have Been Adopted by States and Districts
3. Form Used in Analyzing Standardized Test for Developing Training Objectives

ESEA Title I Evaluation
Regions 8, 9, 10
Technical
Assistance
Centers

November 16, 1981

Ms. Cie Taylor
Utah State University
Logan, UT

Dear Cie:

To find an answer to your question on the tests most frequently used, I contacted David Kaskowitz of RMC Corporation, who is responsible for the national analysis of Title I data. The results of his preliminary analysis show project utilization of tests in this order:

California Achievement Test
SRA
Metropolitan Achievement Test
Gates-MacGinitie
Stanford Achievement Test
Iowa Tests of Basic Skills

Dr. Kaskowitz stressed that the order might change with further analyses and could be quite different when the numbers of students within a project are taken into account. Also, he mentioned that frequencies associated with the first four tests were similar, with a gap between them and the last two.

Do phone again, should you have further questions.

Cordially,

Mary R. Quilling

Mary Quilling
Senior Research Associate
Title I Evaluation
Technical Assistance Center

MQ/pk

cc: Kathy Stewart

BEST COPY AVAILABLE

Northwest Regional Educational Laboratory
300 S.W. Sixth Avenue · Portland, Oregon 97204 · Telephone (503) 295-0214
AN EQUAL OPPORTUNITY EMPLOYER

MIDCONTINENT Region

CAT Districts

St. Louis
Cincinnatti
Omaha
Detroit
Minneapolis
Dayton
DeKalb, IL
Lincoln
Columbus

CTBS Districts

Cleveland

Other Districts

Milwaukee	ITBS
Chicago	ITBS
Kansas City	ITBS
Cleveland	MAT
Indianapolis	ITBS
Chicago Arch Diocese	ITBS
Iowa	ITBS 3-8
Des Moines	ITBS/MAT
Flint	SRA
Toledo	ITBS
Witchita	ITBS

CAT States

None

CTBS States

Wisconsin - 4,8,11
Kentucky - 3,5,7,10

Other States

Iowa - ITBS [illegible]8

DISTRICT TOTALS

CAT	9
CTBS	1
Other	3
ITBS	9

STATE TOTALS

CAT	0
CTBS	2
Other	1

WESTERN Region

CAT Districts

Long Beach
Fresno
Bakersfield
Santa Ana
Seattle
Spokane
Salt Lake City
Phoenix
Tucson
Clark Co., NV
Pasadena

CTBS Districts

Los Angeles
Sacramento
San Diego
San Francisco
Garden Grove
San Jose
Albuquerque
Denver
Jefferson Co., CO
Los Angeles Arch Diocese
Oakland
Tacoma

Other Districts

Clark Co., NV - MAT
Granite, UT - SAT
San Juan, CA - ITBS
Portland - Own Test

CAT States

Washington - 4
Arizona - 1-12

CTBS States

Utah - 4-8 Sample
New Mexico - 5,8,11

Other States

Hawaii - SAT 2,4,6,8,10

DISTRICT TOTALS

CAT	11
CTBS	12
Other	4

STATE TOTALS

CAT	2
CTBS	2
Other	1

SOUTHERN Region

CAT Districts

Memphis
Oklahoma City
Charlotte-Mecklenberg, NC
Akron
Atlanta
Birmingham
Corpus Christi
Mobile
Caddo Parish
El Paso

CTBS Districts

Jefferson Co., KY
Jefferson Parish, LA
Broward Col, FL
Brevard Co., FL
Hillsborough Co., FL
Orange Col, FL
Tallahassee
New Orleans
Charleston, SC
Nashville

Other Districts

Dallas	ITBS
Houston	ITBS
Dade Col, FL	SAT
Miami Diocese	SRA
Orange Diocese	SRA
Tallahassee Diocese	SRA
Tampa Diocese	SRA
Jacksonville Diocese	SRA
New Orleans Diocese	SRA
Jacksonville	SAT
Tulsa	SRA
Pinellas Co.,FL	SRA
Fort Worth	ITBS
Palm Beach	SAT

CAT States

Alabama 1-12
Mississippi - 4,6,8
North Carolina - 3,6,9
Texas - 6 Sample
Oklahoma - 6-9 Sample

CTBS States

South Carolina - 4,7,10

Other States

None

DISTRICT TOTALS

CAT	10
CTBS	10
ITBS	3
SRA	8
SAT	3

STATE TOTALS

CAT	5
CTBS	1
Other	0

EASTERN Region

CAT Districts

New York City
Pittsburg
Pittsburgh Diocese
New Castle Co., DE
Philadelphia
Baltimore
Montgomery Co., MD
Prince George Co., MD
Jersey City
Kanawha Co.

CTBS Districts

Washington, D.C.
Newark, NJ

Other Districts

New York Arch Diocese	SRA
Newark Diocese	SRA
Brooklyn Diocese	SRA
Norfolk	SRA
Richmond	SRA
Rochester	MAT

CAT States

Delaware - 1-8,11
Maryland - 3-5-8

CTBS States

West Virginia - 3-6-9

Other States

Virginia - SRA
Rhode Island - ITBS 4,8

DISTRICT TOTALS

CAT	10
CTBS	2
SRA	5
MAT	1

STATE TOTALS

CAT	2
CTBS	1
Other	2

GRAND TOTALS

DISTRICTS

CAT	39
CTBS	26
	65

MAT-SAT	7
ITBS	13
SRA	14
Own Test	1
	35

STATES

CAT	9
CTBS	6
	15

ITBS	2
SAT	1
SRA	1
	4

MAT	Pr	Reading Compr.	Trider
TEST	LEVEL	SUBTEST.	REVIEWER

For 1 - 4 indicate if the words are written on the test booklet, otherwise, oral will be assumed to be the mode.

I. Difficult vocabulary from directions (individual words).

Sample (written but not used by teacher - only referred to as "find the box..")
first
second
circle

II. Difficult directions (phrases).

answer that best finishes the sentence
mark the space next to the right answer for each question

III. Series of directions (in steps).

1. Find the box at the top of the page
2. Look at the story
3. Read the story to yourself
4. Then read the first question
5. Raise your hand if you know the answer

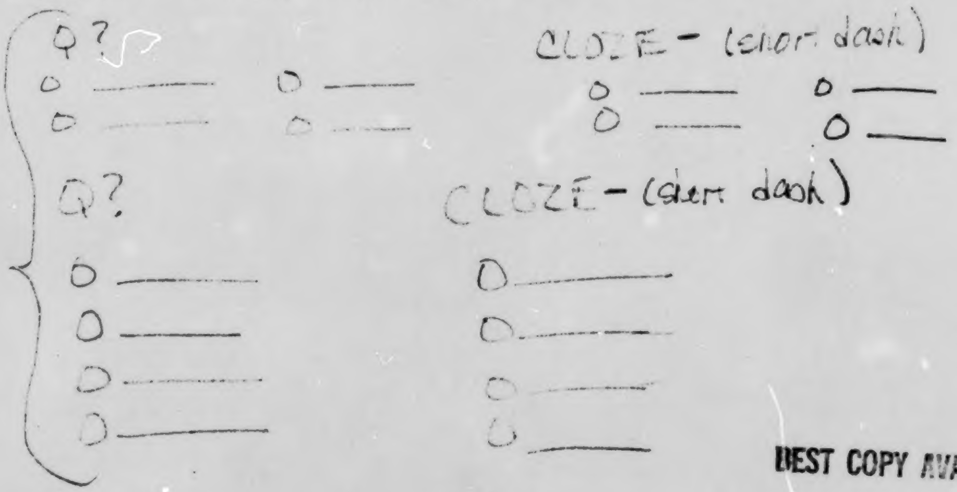
1. Read the stories
2. Answer questions by marking...
3. After finish page, turn to next..
4. fold it back
5. Work on page 3, 4, 5, 6, & 7 until STOP or until I tell you.

IV. New symbols.

○
STOP

V. Examples of different response formats (from test booklet).

lowing
ories
questions
e asked



BEST COPY AVAILABLE

Appendix C

1. Number of Minutes and Items for Each Practice Test in Participating Districts
2. Reading Series Used in Participating Districts Upon Which Practice Tests Were Based
3. Strategies Used to Construct Distractors for Practice Tests
4. Practice Test Directions for Experimental Group I for Test #5

Number of Items and Minutes for Each Subtest of the
Seven CTBS Practice Tests Used in Cache School District

Subtest	PRACTICE TESTS													
	1		2		3		4		5		6		7	
CTBS (1973)	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time
Reading Vocabulary	12	5.52	12	5.52	9	4.14	9	4.14	12	5.52				
Reading Comprehension Sentences			6	5.22	6	5.22	6	5.22	8	6.96				
Reading Comprehension Paragraphs					4	4.67	4	4.67	7	8.17				
CTBS (1981)														
Word Analysis														
A. Consonant Sounds											3	2.25	3	2.25
B. Vowel Sounds (Auditory)											2	1.5	2	1.5
C. Vowel Sounds (Visual)											6	4.5	6	4.5
D. Word Identification											2	1.5	2	1.5
E. Syllables											2	1.5	2	1.5
F. Root Words											2	1.5	2	1.5
G. Compound Words											2	1.5	2	1.5
Vocabulary														
A. Synonyms											4	3.0	4	3.0
B. Sentence Completion											2	1.5	2	1.5
Reading Comprehension											8	8.8	8	8.8
TOTALS	12	5.52	18	10.74	19	14.03	19	14.03	27	20.65	33	27.55	33	27.55

Number of Items and Minutes for Each Subtest of the
Seven ITBS Practice Tests Used in Nebo School District

Subtest	PRACTICE TESTS													
	1		2		3		4		5		6*		7*	
	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time
VCB 1. Picture Identification	9	4.5	6	3.0							6	3.0	6	3.0
VCB 2. Definition			8	4.0	6	3.0					4	2.0	4	2.0
WA 1. Initial Sound (Picture)			4	1.2	4	1.2					2	.6	2	.6
WA 2. Initial Sound (Word)			4	1.2	4	1.2					2	.6	2	.6
WA 3. Final Sound (Picture)					4	1.2	4	1.2			2	.6	2	.6
WA 4. Final Sound (Word)					4	1.2	4	1.2			2	.6	2	.6
WA 5. Sound Substitution					4	1.2	4	1.2			4	1.2	4	1.2
WA 6. Silent Letters							4	1.2	4	1.2	2	.6	2	.6
WA 7. Middle Consonants							4	1.2	4	1.2	2	.6	2	.6
WA 8. Vowel Sounds							4	1.2	4	1.2	4	1.2	4	1.2
WA 9. Long/Short Vowels							4	1.2	4	1.2	4	1.2	4	1.2
WA 10. Endings									4	1.2	2	.6	2	.6
WA 11. Compound Words									4	1.2	2	.6	2	.6
Picture Description					6	3.0	6	3.0	9	4.5	12	6.0	12	6.0
Sentence Understanding							4	1.6	8	3.2	10	4.0	10	4.0
Stories									8	4.0	14	7.0	14	7.0
TOTALS	9	4.5	22	9.4	32	12.0	38	13.0	49	18.9	74	30.4	74	30.4
LIMIT		5		10		15		15		20		30		30

*Use sample items only as indicated in the test.

Number of Items and Minutes for Each Subtest of the Seven SAT Practice Tests Used in Granite School District

Subtest	PRACTICE TESTS													
	1		2		3		4		5		6		7	
	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time
Vocabulary	9	4.86	7	3.78					7	3.78	13	7.02	13	7.02
Reading A (Picture Identification)			12	5.40	12	5.40			12	5.40	12	5.40	12	5.40
Reading B (Sentence Completion)					10	5.20	10	5.20	8	4.16	13	6.76	13	6.76
Word Study Skills A (Word Identification)					13	4.29	13	4.29	8	2.64	13	4.29	13	4.29
Word Study Skills B (Sound Discrimination)							12	5.16	12	5.16	12	5.16	12	5.16
TOTALS	9	4.86	19	9.18	35	14.89	35	14.65	47	21.14	63	28.63	63	28.63

Number of Items and Minutes for Each Subtest of the
Seven MAT Practice Tests Used in Logan School District

Subtest	PRACTICE TESTS													
	1		2		3		4		5		6		7	
	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time	Items	Time
Word Knowledge A (Picture Identification)	15	5.25	15	5.25					6	2.10	15	5.25	15	5.25
Word Knowledge B (Definition)			12	6.24	12	6.24			8	4.16	12	6.24	12	6.24
Word Analysis					12	5.16	12	5.16	12	5.16	12	5.16	12	5.16
Reading Sentences					7	3.78	7	3.78	4	2.16	7	3.78	7	3.78
Reading Stories							7	5.18	9	6.66	14	10.36	14	10.36
TOTALS	15	5.25	27	11.49	31	15.18	26	14.12	39	20.24	60	30.79	60	30.79

EXPERIMENTAL GROUP I
CLASSROOM TEXT INFORMATION

District	School	Teacher	Series	Level	Title
Cache	Wellsville	V. Jenkins	Holt	7	A Place For Me
				9	People Need People
		L. Murray	Holt	7	A Place For Me
				9	People Need People
		C. Nielsen	Holt	7	A Place For Me
				9	People Need People
Granite	Hillsdale	P. Jensen	Distar	II	Fast Cycle
				II	Book B
			Houghton-Mifflin	6	Secrets
		P. Kane	Distar	II	Fast Cycle
				II	Book B
			Houghton-Mifflin	6	Secrets
		G. Kunz	Distar	II	Fast Cycle
				II	Book C
			Houghton-Mifflin	6	Secrets
		S. Waldram	Distar	II	Fast Cycle
				II	Book B
			Houghton-Mifflin	6	Secrets
	Lincoln	E. Archer	Ginn	3,5,6	A Duck Is A Duck May I Come In One to Grow On
		A. Norris	Ginn	5,6,7	The Dog Next Door
	Redwood	B. J. Crockett	Distar	II	Fast Cycle
				II	Book C
			Ginn	6	One to Grow On

District	School	Teacher	Series	Level	Title
Granite	Redwood	V. Latham	Distar	II	Book A
				II	Book C
			Ginn	6	One to Grow On
	West Kearns	E. Banks	Distar	I	Book C
				II	Book A
			Ginn	7	The Dog Next Door
		C. Borden	Distar	I	Book B
				II	Book A
			Ginn	6	One to Grow On
		V. Gomez	Distar	I	Book A
				II	Book A
			Ginn	7	The Dog Next Door
		S. Green	Ginn	7	The Dog Next Door
				8	How It Is Nowadays
		L. Lobb	Distar	II	Book A
			Ginn	8	How It Is Nowadays
		F. Martin	Distar	I	Book C
				II	Book B
			Ginn	7	The Dog Next Door
Nebo	Santaquin	M. Anthony	Harcourt Brace Jovanovich	5	Together We Go
				6	A World of Surprises
		M. Willis	Harcourt Brace Jovanovich	5	Together We Go
				6	A World of Surprises
	Westside	A. Burbidge	Harcourt Brace Jovanovich	5	Together We Go
				6	A World of Surprises
				7	People and Places
		M. Payne	Harcourt Brace Jovanovich	5	Together We Go
				6	A World of Surprises
				7	People and Places

BEST COPY AVAILABLE

EXPERIMENTAL GROUP II
CLASSROOM TEXT INFORMATION

District	School	Teacher	Series	Level	Title
Cache	Park	E. Taggart	Holt	7	A Place For Me
				9	People Need People
		L. Talbot	Holt	9	People Need People
				11	?
	Lewiston	D. Mieure	McMillan	7A	On Wings of Words
			Holt	9	People Need People
		M. Schenever	McMillan	6	Worlds of Wonder
				7	Lands of Pleasure
				7A	On Wings of Words
Granite	South Kearns	M. Franco	Distar	II	Book A
			Ginn	7	The Dog Next Door
				8	How It Is Nowadays
		G. Madsen	Distar	II	Book A
			Ginn	7	The Dog Next Door
				8	How It Is Nowadays
		E. Zagarella	Distar	II	Book A
				II	Book B
			Ginn	8	How It Is Nowadays
	Stansbury	B. Hunt	Distar	II	Fast Cycle
			Houghton-Mifflin	6	Secrets
				7	Rewards
		M. Miller	Houghton-Mifflin	4	Rainbows
				7	Rewards
				8	Panorama
		L. Sorensen	Distar	II	Book B
			Houghton-Mifflin	7	Rewards
				8	Panorama

District	School	Teacher	Series	Level	Title
Granite	Stansbury	O. Wallace	Distar	II	Fast Cycle
				II	Book A
			Houghton-Mifflin	7	Rewards
	Western Hills	B. Cannon	Distar	II	Fast Cycle
				II	Book C
			Houghton-Mifflin	?	Spinners
		J. Eber	Distar	II	Fast Cycle
				II	Book B
			Houghton-Mifflin	?	Spinners/Towers/Skylights
		J. Schmidt	Distar	II	Fast Cycle
				II	Book D
			Houghton-Mifflin	? (2.5)	2.5 (?)
		D. Tanner	Distar	II	Fast Cycle
				II	Book C
			Houghton-Mifflin	?	Towers
		M. Turner	Distar	II	Fast Cycle
				II	Book C
			Houghton-Mifflin	?	Towers
Nebo	Goshen	R. Boyack	Lynn & Bacon	Special Primer	At Home and Away
			Scott Foresman	1-10	Calico Capers
			Scott Foresman	2-1	Daisy Days
		L. Neff	Lynn & Bacon	Special Primer	At Home and Away
			Scott Foresman	1-10	Calico Capers
			Scott Foresman	2-1	Daisy Days
	Wilson	D. Altenberg	Houghton-Mifflin	E	Honeycomb
			Houghton-Mifflin	F	Clover Leaf
			Houghton-Mifflin	G	Sunburst
		M. Anderson	Houghton-Mifflin	E	Honeycomb
			Houghton-Mifflin	F	Clover Leaf
			Houghton-Mifflin	G	Sunburst

SAT
TEACHER'S KEY
PRACTICE TEST # 5

Vocabulary	Reading A	Reading B	Word Study A	Word Study B
1. c	1. b	1. a	1. a	1. b
2. b	2. a	2. b	2. b	2. b
3. b	3. c	3. d	3. c	3. a
4. a	4. a	4. b	4. b	4. c
5. c	5. a	5. d	5. a	5. a
6. a	6. b	6. a	6. ____	6. c
7. ____	7. b	7. ____	7. ____	7. a
8. ____	8. b	8. ____	8. ____	8. b
9. ____	9. c	9. ____	9. ____	9. a
10. ____	10. ____	10. ____	10. ____	10. b
11. ____	11. ____	11. ____	11. ____	11. a
12. ____	12. ____	12. ____	12. ____	12. ____
13. ____		13. ____	13. ____	

SAT
VOCABULARY

	DO	SAY	GROUP RESPONSE
1	Demonstrate. Check fingers.	Turn to page 1. Put your finger on page number 1. This is a Vocabulary test. This vocabulary test will show how many words you know.	What is this? What will this vocabulary test show?
2		Put your finger on the sample girl , boy , boot .	Read the words out loud with me.
3		First, I will read part of a sentence. Then I'll read three words. You will have to find which of the three words completes the sentence.	What will I read first? Then what will I read?
4		Let's try the sample. Listen to the sentence. A young man is a ______. girl , boy , boot .	Try each word to find which word completes the sentence. Try girl . A young man is a , girl . Is it right? Try boy . A young man is a , boy . Is it right? Try boot . A young man is a , boot . Is it right?
5		Which word completes the sentence? Yes, boy . You can see the space under boy has been marked.	
6		We will do all the items on this page the same way.	
7	Check fingers.	Finger on item number 1. Listen for the word that best completes the sentence.	Read the words to yourself.

VOCABULARY (continued)

	DO	SAY	GROUP RESPONSE
8	Make sure all students made a mark.	1. You live at ______, school, dinner, home. You live at ______, school, dinner, home.	Mark the answer space.
9	Wait 10 seconds between items. Repeat #8 for these items. Say each sentence twice.	2. A type of fruit is an ______, ape, apple, acorn. 3. A short sleep is a ______, nut, nap, snap. 4. We get wool from ______, sheep, sleep, fur. 5. A smile is a ______, snarl, grim, grin. 6. Seven days make a ______, week, weed, weak. 7. ______, ______, ______, ______. 8. ______, ______, ______, ______. 9. ______, ______, ______, ______. 10. ______, ______, ______, ______. 11. ______, ______, ______, ______. 12. ______, ______, ______, ______. 13. ______, ______, ______, ______.	

SAT
READING A

	DO	SAY	GROUP RESPONSE
1	Check fingers.	Turn to page 2. Put your finger on page number 2. This is a Reading test.	What is this?
2		This reading test asks you to find words that go with the picture.	This reading test asks you to find words that go with what?
3		Look at the sample and put your finger on the picture of hut. There are three lines under the picture. There is one right answer in each line and there are three words to choose from.	How many words in each line? How many right answers in each line?
4	Check fingers.	Finger on A. hut, hum, room. Which word tells about the picture? Yes, see the space under hut has been marked.	Read the words with me.
5	Check fingers. Check marks.	Finger on B. loft, house, hound. Which word tells about the picture? Yes, mark the space under the word house.	Read the words with me.
6	Check fingers. Check marks.	Finger on C. fix, hoe, home. Now mark the space under the word that tells about the picture. You should have marked under home.	Read the words with me.
7		Now you will do the rest of the items on this page just like the sample. When you get finished, go back and check your work.	What do you do when you're finished?

SAT
READING A (continued)

	DO	SAY	GROUP RESPONSE
8	Record time. Start __ : __ Time 5 :00 Stop __ : __	Finger on item number 1. Go.	
9		Stop.	

SAT

READING B

	DO	SAY	GROUP RESPONSE
1	Check fingers.	Turn to page 3. Put your finger on page number 3. This is a Reading test. You will read sentences and stories. Then you will answer questions about the sentences and stories.	What will you read?
2	Check fingers.	Put your finger on the sample. We saw a bright colored bird at the zoo. The color of the bird might be ______.	Read the sentences with me.
3	Check fingers.	Now you will pick a word to finish the sentences. Finger on A. red, pale, gray, dull.	Read the words.
4		Which words tell the color of the bird?	
5		Yes, the word red has been marked.	
6	Check marks.	The next sentence reads . . . The bird was in the ______. Now, we will read the words beside B. soup, hat, zoo, animal. Which word tells where the bird was? Mark the word you think finishes the sentence. You should have marked zoo.	Read the sentence with me.
7		Now you will do the rest by yourself. When you get finished, go back and check your work.	What do you do when you're finished?
8	Record time. Start ___ : ___ Time 4 : 00 Stop ___ : ___	Finger on item number 1. Go.	
9		STOP.	

SAT
WORD STUDY SKILLS - A

	DO	SAY	GROUP RESPONSE
1	Check fingers.	Turn to page 4. Put your finger on page number 4. For this test you will find a word that I say.	
2	Check fingers.	Put your finger on the sample, line A. paste, past, patch. Find the word paste. See the space below paste has been marked.	Read the words out loud with me.
3	Check fingers. Check marks.	Finger on line B. sick, six, sill. Mark the word six, six. Is six the first, middle, or last word? Yes, the middle word.	Read the words with me.
4	Check fingers. Check marks.	Finger on line C. pill, peg, pig. Mark the word pig, pig. You should have marked the last word.	Read the words with me.
5		Now we will do all of the items on this page together. Remember to mark only the word I say.	
6	Check fingers.	Finger on item 1. Mark wept, wept.	Read the words to yourself.
7.	Wait 10 seconds between items.	Item 2. Mark reach, reach.	
8	Repeat #7 for these items. Say each word twice.	3. wheel 7. ______ 11. ______ 4. little 8. ______ 12. ______ 5. gloat 9. ______ 13. ______ 6. ______ 10. ______ 14. ______	

SAT

WORD STUDY SKILLS - B

	DO	SAY	GROUP RESPONSE
1	Check fingers.	Turn to page 5. Put your finger on page number 5. This is a test on the sounds that letters make.	
2	Check fingers.	Put your finger on the sample. Read the first word to yourself. The SS has a line under it. Listen for the sound of the underlined letter(s).	Now read it out loud. What letter(s) is underlined? What sound is underlined?
3	Point to 3 words.	Now you will find the underlined sound in one of these three words. Read the three words to yourself.	Now read the words out loud.
4		Which word has the underlined sound? Is it city, goal, or rail?	Say the underlined sound again. Remember to listen for the underlined sound, not letter.
5		Yes, city. The space under city has been marked.	
6		Now you will do the rest by yourself. When you get finished, go back and check your work.	What do you do when you're finished?
7	Record time. Start __:__ Time 5:00 Stop __:__	Finger on item number 1. Go.	
8		STOP.	

TEST 1.

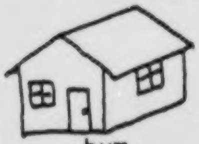
Vocabulary

	girl	boy	boot
SAMPLE:	a ○	b ●	c ○
	school	dinner	home
1	a ○	b ○	c ○
	ape	apple	acorn
2	a ○	b ○	c ○
	nut	nap	snap
3	a ○	b ○	c ○
	sheep	sleep	fur
4	a ○	b ○	c ○
	snarl	grim	grin
5	a ○	b ○	c ○
	week	weed	weak
6	a ○	b ○	c ○

STOP

TEST 2.
Reading: Part A

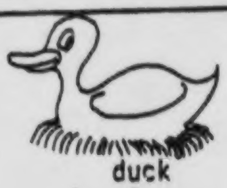
.MPLE:



a a. hut ● b. hum ○ c. room ○

b a. loft ○ b. house ○ c. hound ○

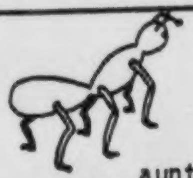
c a. fix ○ b. hoe ○ c. home ○



1 a. dock ○ b. duck ○ c. dish ○

2 a. nest ○ b. needle ○ c. home ○

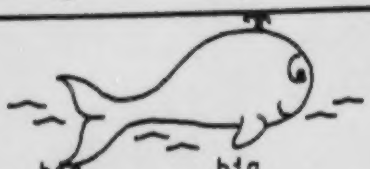
a. crane ○ b. flew ○ c. bird ○



4 a. ant ○ b. aunt ○ c. rob ○

5 a. bug ○ b. brush ○ c. bait ○

6 a. step ○ b. pest ○ c. ladder ○



7 a. bag ○ b. big ○ c. crash ○

8 a. zoo ○ b. animal ○ c. acorn ○

9 a. white ○ b. wish ○ c. whale ○

STOP

TEST 2.

Reading: Part B

SAMPLE:

We saw a bright colored bird at the zoo.
The color of the bird might be

A.	red	pale	gray	dull
	a●	b○	c○	d○

The bird was in the

B.	soup	hat	zoo	animal
	a○	b○	c○	d○

He made a statue from a tree.
The statue was made of

1.	wood	metal	glass	clay.
	a○	b○	c○	d○

She put the basket on her head.
She wanted to wear a

2.	dish	hat	dress	shoe.
	a○	b○	c○	d○

Mary tries very hard. She always
does her

3.	fast	sleep	bean	best
	a○	b○	c○	d○

to do a good

4.	fight	job	play	quick.
	a○	b○	c○	d○

Harry likes birds. He built a

5.	river	truck	cabin	cage
	a○	b○	c○	d○

so he could keep a

6.	pigeon	mouse	dog	goat.
	a○	b○	c○	d○

TEST 3.
Word Study Skills: Part A

SAMPLE:

a	paste a ●	past b ○	patch c ○
b	sick a ○	six b ○	sill c ○
c	pill a ○	peg b ○	pig c ○

1	wept a ○	wait b ○	weep c ○
2	read a ○	reach b ○	real c ○
3	when a ○	wheat b ○	wheel c ○
4	litter a ○	little b ○	liter c ○
5	gloat a ○	goat b ○	glow c ○

STOP

TEST 3.

Word Study Skills: Part B

AMPLE:

grass	city a ●	goal b ○	rail c ○

1	home	house a ○	coal b ○	keep c ○
2	face	air a ○	craft b ○	ran c ○
3	paper	wait a ○	ant b ○	pop c ○
4	trumpet	out a ○	trick b ○	cup c ○
5	crust	rake a ○	city b ○	boil c ○

6	raft	ram a ○	flip b ○	flat c ○
7	greet	sleet a ○	tip b ○	grow c ○
8	sloppy	go a ○	hot b ○	slip c ○
9	grouch	witch a ○	truck b ○	dish c ○
10	vine	veil a ○	slide b ○	foil c ○
11	glue	zoo a ○	up b ○	build c ○

STOP

Test Item Construction Strategies

I. Vocabulary (Word Knowledge) Distractors
  A. Picture Identification (match word with picture--nouns and verbs)
    1. Initial Sounds (flower-flame, cup-cut, whisper-whistle-whisker)
    2. Final Sounds (tear-near, goat-boat)
    3. Word Appearance (bitter-butter-batter, number-notice, captain-capture)
    4. Similar Sounds (sheep-sleep-geese, rug-rag, six-sick)
    5. Similar Definition (hood-mask-helmet)
    6. Similar Spelling (rocker-rocket, broad-board)
    7. Related Words (pink-flower, cage-keep, drown-drift)
  B. Simple Definition (match word with short definition--sometimes opposites)
    1. Similar Forms (only-once, quiet-quit, delay-depart, confess-confuse)
    2. Opposites (going-coming, alert-asleep, remain-leave)
    3. Related Parts (stem-root-core, flock-nest)
    4. Related Family of Words (grass-tree, dog-bird)
    5. Incorrect Logic (invited means liked, adult means healthy)
    6. Similar Sounds (light-bright, might-right)

II. Word Analysis Distractors
  A. Similar Appearance (love-live, pear-peal-peat)
  B. Reversal (evil-live)
  C. Similar Sounds (stuffy-fluffy-puffy)
  D. Prefixes (upset-inset-reset)
  E. Spellings (weigh-way, sear-seer, whether-weather, leaf-leave)

III. Reading Comprehension
  A. Sentences
    1. Visual Discrimination (involving pictures of sentences)
    2. Understanding (action of noun in sentence)
    3. Common Sense (Fan is used to make air warmer?)
    4. General Knowledge (Moon means it is night.)
    5. Vocabulary (bubble, elephant, corner, flame)
    6. Logic (double the amount is twice as much)
    7. Inference (how would you feel)
    8. Relationships (brother, sister)
  B. Stories
    1. Summary (title of story)
    2. Sorting out details (who, what, where, when)
    3. Inference (How did Sue feel?)
    4. Common Sense (Do you get wet when bathing?)
    5. General Knowledge (Is Sunday before Monday?)
    6. Judgement (Are Giants good or bad?)
    7. Vocabulary (What word in the story means _____?)
    8. Conclusions

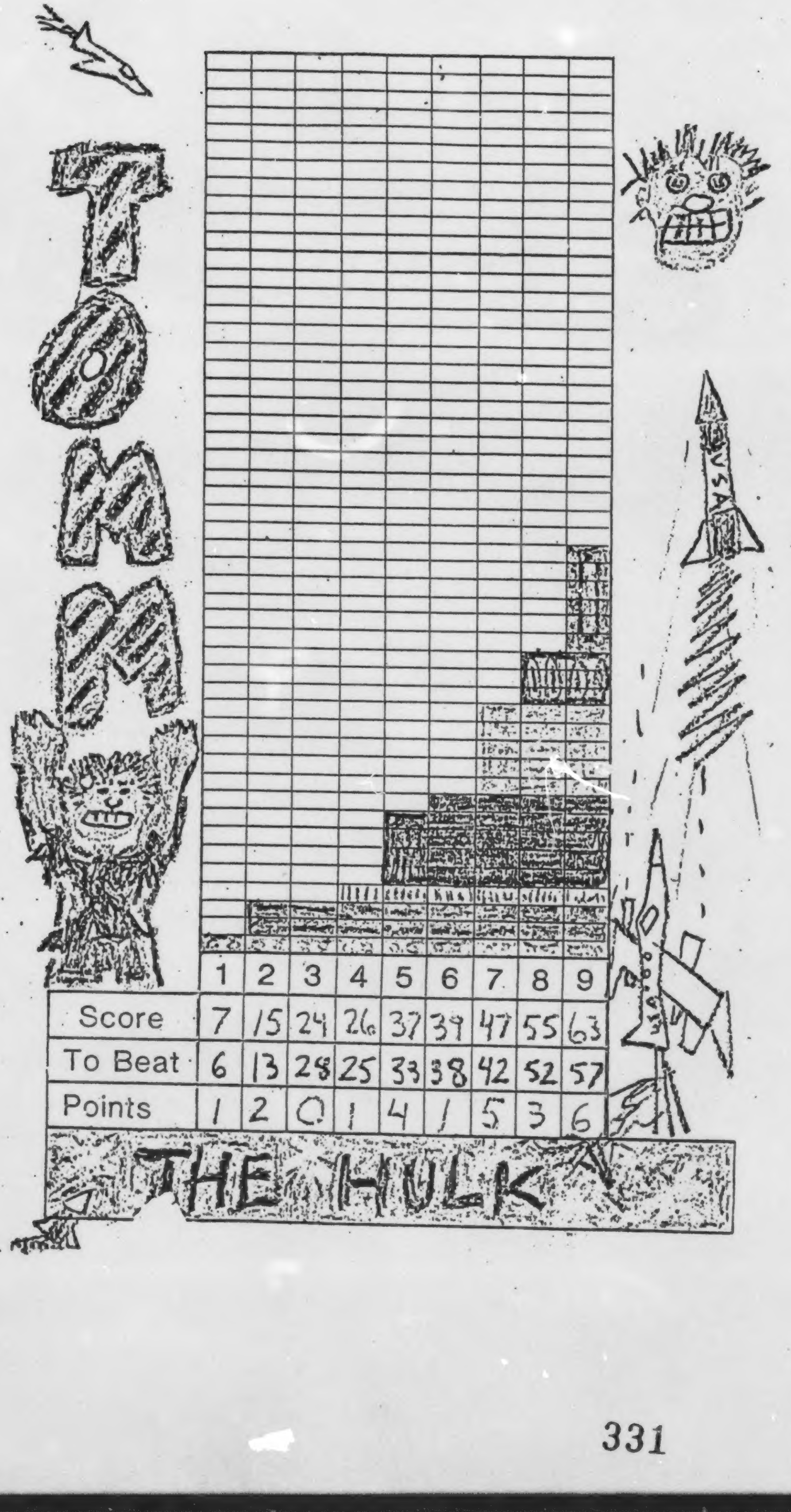
Appendix D

Materials Related to Reinforcement Procedures

1. Sample Chart Used in Reinforcement Component
2. Example of a Completed Chart Used in Reinforcement Component

	1	2	3	4	5	6	7	8	9	10	11	12
Score												
To Beat												
Points												

Note: This chart has been reduced for binding purposes. The actual chart measures 8-1/2" X 14".


TOMMY
USA
THE HULK

	1	2	3	4	5	6	7	8	9
Score	7	15	24	26	37	39	47	55	63
To Beat	6	13	28	25	33	38	42	52	57
Points	1	2	0	1	4	1	5	3	6

Appendix E

Materials Related to Sample Selection and Description

1. Sample Letter Sent to Principals to be Used to Inform Teachers About the Project
2. Letter Sent to Inform Teachers of Assignment to Experimental Group I

SAMPLE LETTER TO BE SENT TO PRINCIPALS

The Utah State Office of Education has been awarded a contract from the U.S. Department of Education to develop, implement, and evaluate the effectiveness of a project to improve the quality of data obtained from standardized achievement tests. I have reviewed the description of the project carefully and am convinced that our district would gain much by participating. Most of the work for the project will be carried out by researchers at Utah State University under the direction of Dr. Karl White.

The purpose of the study is to investigate the effects on standardized test performance of the following variables:

1. Training students in test-taking skills.

2. Reinforcing students for trying their best on standardized tests.

3. Familiarizing students with the format of the particular standardized test used in their district.

A related project was conducted during the last two years by the State Office of Education in conjunction with the researchers at USU. The results of this previous project indicated that the above variables have substantial effect on the results of standardized test performance of elementary school children. The current project will focus on standardized reading achievement tests for second graders. The findings of the previous project will be used to develop and evaluate a number of training packages and procedures. If the project is successful, we will be able to be more confident that the results of our standardized tests are an accurate reflection of what students do or do not know.

Our district has agreed to participate in the research and has suggested that your school (among others) be involved. Second grade teachers from each of the participating schools will be asked to participate. Once it is determined which teachers are willing to participate, the research procedures require that they be randomly assigned to either an experimental or control group. Those teachers assigned to the experimental group will be given training in appropriate test administration techniques and will be trained in how to assist in teaching their students appropriate test-taking and motivational techniques. Those assigned to the control group will receive no training. Data will be collected from all classrooms, but this will require almost no time from the teacher.

Experimental group teachers will need to attend two workshops, one in early September, the other in early spring, to acquaint them with the research rationale and procedures. Since the first workshop will last a whole day and will be held on a Saturday, teachers will be paid an honorarium for attending. The research staff from Utah State University will work directly with the teachers in the experimental group to assist them in implementing the project. In addition, the research team members will be in monthly contact by telephone to offer any other assistance the teachers may find helpful.

BEST COPY AVAILABLE

I think this project will provide valuable training to our students and teachers regarding standardized achievement test administration. Furthermore, the study is important for developing methods which will increase the validity of achievement test scores and provide a more accurate reflection of what students do or do not know. Therefore, I encourage you to support and participate in the project to the extent necessary.

Enclosed in this packet are letters to be sent to the following second grade teachers in your school:

1) ______________________________
2) ______________________________
3) ______________________________
4) ______________________________
5) ______________________________

If you agree with me that our district should participate in this study, please sign each letter and forward them as soon as possible to each of the teachers. Members of the project team will then be contacting each of these teachers by phone to determine which ones are able and willing to participate (I anticipate that a few teachers in the district will have legitimate reasons why they can't participate, but hope that there will not be many). Once we determine which teachers are able to participate, they will be randomly assigned to one of the experimental or control groups and the project will proceed.

I would like to thank you in advance for whatever time and attention you are able to devote to this research. If you have any questions or for some reason think it would be better if your school did not participate, please contact me as soon as possible.

Sincerely,

UTAH STATE UNIVERSITY · LOGAN, UTAH 84322

801-750-1981

UNIVERSITY AFFILIATED
EXCEPTIONAL CHILD CENTER
UMC 68

September 8, 1981

Salt Lake City, UT 84106

Dear Ms.

I am writing concerning the research project being conducted by the Utah State Office of Education in conjunction with Utah State University. As explained to you on the phone, it was necessary to randomly assign those teachers who were willing to participate in the project to various experimental and control groups in order to investigate the effects on standardized test performance of training students in test-wiseness skills and reinforcing students for trying their best. In consultation with your district staff, your school was assigned to the experimental group which will be implementing procedures for student training, reinforcement, and teacher training in test administration.

A workshop will be held on September 12 starting at 9:00 a.m. at The Sirloin Stockade Restaurant located at 972 East 7200 South in Salt Lake City. Since the workshop will take place on Saturday, you will be paid an honorarium of $50 for attending. Lunch will be provided and you should plan on being finished by about 4:00 p.m. An agenda for the workshop is enclosed.

To help us in getting the project off to a good start, there are a number of things you need to bring to the workshop. These are listed below:

1. Reading Series Materials. As a part of the project, we will be preparing practice tests for you to give to your students during the year. These practice tests will be based on the Reading Series you are using in your class. Therefore, please bring with you a copy of (a) the Teachers Manual, (b) the student text, and (c) the student workbook. If your class is using multiple levels, please bring all levels with you. Also, we will need to use these materials regularly during the year, so bring copies that we can keep (if all teachers in the district use the same materials, we will only need one copy of each level, but we can arrange that at the workshop).

2. List of class members.

3. Results of WHAT? To get you into the swing of the workshop, we have attached an abbreviated copy of the WorksHop Achievement Test which will serve as your name tag for the workshop. Please complete the test and bring it with you as per the instructions.

It is very important that you attend this workshop, since it will explain and demonstrate all of the procedures and materials which will be used during the project. If something comes up that makes it impossible for you to attend, please contact me as soon as possible at (801) 750-2003.

On behalf of the State Office of Education and your school district administration, I would like to thank you for your willingness to participate. I know that as a teacher, you already have more to do than can reasonably be expected, and your willingness to add another concern to your daily affairs (even though this project will take relatively little time) is much appreciated. We believe that the results of this project will do much to assist us in understanding and making more accurate the results of standardized achievement tests.

Sincerely,

Karl R. White

Karl R. White, Ph.D.
Director, Planning & Evaluation

KRW:mmt

Enclosure

Appendix F

Materials Related to Implementation of Training Materials

1. Filmstrip Evaluation Form
2. Project Evaluation Form

FILMSTRIP EVALUATION

(Please send this to USU with your next Practice Test)

School ________ District ________ Teacher ________

Filmstrip # ____ Filmstrip shown on ________ at ________

Date Time

I. Please rate the following statements according to this scale:

Filmstrip

	STRONGLY AGREE	AGREE	DISAGREE	STRONGLY DISAGREE
1. The length was appropriate.	1	2	3	4
2. The story line was entertaining to the students.	1	2	3	4
3. The content addressed skills the students need to learn.	1	2	3	4
4. The figures and printing on the filmstrip were clear.	1	2	3	4
5. The dialogue was audible.	1	2	3	4
6. The filmstrip turner was able to move with the narrated page.	1	2	3	4

Teacher Involvement

7. The teacher was properly cued to stop the tape.	1	2	3	4
8. The amount of Owl/teacher interaction was appropriate.	1	2	3	4
9. The tasks required of the teacher were easy to accomplish and defined clearly.	1	2	3	4

Student Materials

10. The student practice was sufficient for students to apply the concepts they learned through the filmstrip.	1	2	3	4
11. The practice exercises were of the appropriate difficulty level.	1	2	3	4

II. Please answer the following questions.

1. Have the students applied their test-taking skills to other subjects?
   Yes No
   In what way? ________________________________
   ____

2. How long did it take you to prepare to teach this filmstrip? ________

3. Were there any concepts presented in the filmstrip that were not learned by your students? Yes No
   Describe ________________________________
   ____

4. Were you the teacher for the filmstrip? Yes No

5. Did you use the pictures that accompany the filmstrip? Yes No
   How? ________________________________
   ____

6. If you have any additional comments, please write them on the back of this form.

(Name)

TEST-TAKING SKILLS PROJECT EVALUATION FORM

INSTRUCTIONS: Listed below are statements about each component of the project to which we would like you to respond. Please circle the number that indicates how you feel about each item. To save your time, we have not left space for you to write open-ended comments. Instead, a member of our staff will soon contact you by phone for you to summarize your comments about the best and worst aspects of each project component and how the project could be improved. After the phone call, please return this form in the enclosed envelope. Please be as candid and specific as possible, so we will know which parts are good and which parts need to be improved. Thank you.

FILMSTRIPS	Strongly Agree		Neutral		Strongly Disagree
1. Instructions for teachers were complete and easy to follow	1	2	3	4	5
2. The filmstrips were easy to implement in the classroom	1	2	3	4	5
3. The concepts taught in the filmstrips were important for students to learn . .	1	2	3	4	5
4. The filmstrips taught the concepts adequately	1	2	3	4	5
5. The students enjoyed the filmstrips. . .	1	2	3	4	5
6. I plan to use the filmstrips in future classes	1	2	3	4	5
7. The filmstrips were worth the time and effort required.	1	2	3	4	5

PRACTICE TESTS					
8. Directions to students were complete and easy to follow	1	2	3	4	5
9. Tests were easy to implement in the classroom.	1	2	3	4	5
10. The test items were appropriate in terms of content and difficulty.	1	2	3	4	5
11. The tests adequately prepared the students for standardized testing. . . .	1	2	3	4	5
12. I plan to use the practice tests in the future	1	2	3	4	5
13. Students enjoyed taking the practice tests	1	2	3	4	5
14. The practice tests were worth the time and effort required	1	2	3	4	5

(over)

CONTACT AND COMMUNICATION	Strongly Agree		Neutral		Strongly Disagree
15. The USU contact person kept me well informed	1	2	3	4	5
16. I was able to reach my USU contact person and felt comfortable in doing so.	1	2	3	4	5
17. My needs were responded to in a reasonable amount of time.	1	2	3	4	5
18. The contact person listened and responded to my feedback	1	2	3	4	5
DATA COLLECTION					
19. The observation during testing was non-disruptive	1	2	3	4	5
20. I would not mind having observers again in a similar project	1	2	3	4	5
21. Students enjoyed responding to the student attitude measures on Friday.	1	2	3	4	5
GENERAL IMPRESSIONS					
22. The requirements for participation in the study were clearly outlined	1	2	3	4	5
23. The benefits were worth the investment of time	1	2	3	4	5
24. The project was enjoyable for students	1	2	3	4	5
25. The project benefited students' test-taking ability.	1	2	3	4	5
26. The project enhanced students' performance in other areas	1	2	3	4	5
27. The project was realistic in scope	1	2	3	4	5
28. I am glad that I participated.	1	2	3	4	5
29. The fall workshop adequately prepared me for the tasks expected.	1	2	3	4	5
30. Taking tests was less anxiety-provoking for students because of the project.	1	2	3	4	5
REINFORCEMENT					
31. The reinforcement procedures were easy for students to understand.	1	2	3	4	5
32. The reinforcement procedures were easy for the teacher to implement.	1	2	3	4	5
33. Students worked hard to earn more than their "to beat" score on the test.	1	2	3	4	5
34. Students enjoyed the reinforcement procedures	1	2	3	4	5
35. I plan to use the procedures for reinforcement in the future.	1	2	3	4	5
SPRING WORKSHOP					
36. Workshop materials were clear and helpful.	1	2	3	4	5
37. Workshop was appropriate in length	1	2	3	4	5
38. Information gained from the workshop(s) was worth the amount of time required.	1	2	3	4	5
39. As a result of the workshop, I was a better test administrator.	1	2	3	4	5

Appendix G

Materials Related to Instrumentation

1. Instrument (with Percentage of Teachers Observed Doing Each Alternative Broken Down by Group) Used to Collect Data on Quality of Test Administration
2. Instrumentation and Explanations Used to Collect Data on Student and Teacher On-Task Behavior During Standardized Testing
3. Observer Training Outline and Schedule
4. Procedures for Observers to Collect the Data for 6 Measures: Quality of Test Administration, Teacher and Student On-Task Behavior, Student and Teacher Attitude, and Student Test-Wiseness
5. Schedules for Teachers and Observers for Classroom Visits During Testing Week
6. Instrument (with Means and Standard Deviations for Each Item and Subscale Broken Down by Group) Used to Collect Data on Teacher's Attitude Toward Standardized Tests
7. Directions for Administering Student Attitude and Student Test-Wiseness Forms
8. Instrument (with Means and Standard Deviations for Each Item Broken Down by Group) Used to Collect Data on Student's Attitude Toward Standardized Tests
9. Instrument (with Percentage of Respondents Selecting Each Option Broken Down by Group) Used to Collect Data on Student's Test-Wiseness Skills

OBSERVATION FOR STANDARDIZED ACHIEVEMENT TESTING

Teacher ____________ School ____________ Date ________ Time ______ SD ___ TD ___ Observer ___________ Partner ________

	STUDENT #1				STUDENT #2				STUDENT #3				STUDENT #4				STUDENT #5				TEACHER			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
a																								
b																								
c																								
d																								
e																								
f																								
g																								
h																								
i																								
j																								
k																								
l																								
m																								
n																								
o																								
p																								
q																								

STUDENT #1	STUDENT #2	STUDENT #3	STUDENT #4	STUDENT #5	TEACHER
TOTAL ONTASK ____	TOTAL ONTASK ____	TOTAL ONTASK ____	TOTAL ONTASK ____	TOTAL ONTASK ____	TOTAL ONTASK ____
% ONTASK ______	% ONTASK ______	% ONTASK ______	% ONTASK ______	% ONTASK ______	% ONTASK ______

Unusual Circumstances:

CODE:

1	Ontask (for entire interval)
-	Probably ontask
0	Offtask (for part of interval)

☐ Beginning of test time

☐ End of timed test

☐ No record made (Explain in NOTES section)

Directions: Record 4 intervals on one student before observing nex student. Observe 5 students and one teacher for a total of 24 intervals before repeating sequence.

Observer Training

1. Introduction to tests
   a. group administered standardized achievement tests/show example
   b. machine scorable/multiple choice format/no separate answer sheet
   c. tests cover reading and math
   d. only observe reading subtests
   e. both TD and SD
      - TD example - word study
      - SD example - timed test/vocabulary, comprehension
   f. observe both teacher and students

2. Types of observations
   a. two types: checklist and interval recording
   b. training is important to clearly define the parameters
      - to increase reliability/consistency
   c. not feasible to record all behaviors - no way to summarize
   d. reduce to categories - numbers - data analysis
   e. work in pairs

3. Quality of Test Administration Checklist
   a. go over heading
   b. Class Environment
      TAPE - Stop at "hurry boys"
   c. Student Preparation - Remind Students
      TAPE - stop at "I'm going to give . . . ."
   d. Positive Atmosphere and Reading Directions
      TAPE - Stop at "Stop Tape"
   e. End of test - after test
      Fill in checklist

4. Teacher On-Task
   a. go over definition
   b. watch for teacher on-task
      TAPE - Stop when teacher moves over

5. Student On-Task
   a. go over definition
   b. watch for student on-task
      TAPE - Stop at "stop tape"

6. Observation form

   a. go over items on form
   b. play tape - listen to intervals
   c. explain use of entire interval
   d. practice on tape - start at 515 (timed) - no teacher
   e. check standard
   f. practice on tape - start at 421 (timed) with teacher

7. Complete Rehearsal

   a. organize materials
   b. go over observation procedures
   c. leave room and return to set up
   d. start checklist at 277
   e. start interval at 340 (teacher directed)
      direction giving)
   f. continue teacher directed at 376
   g. complete checklist at 402 (Stop interval)(Channel 1)

8. Schedule

   a. district schedule
   b. consultant forms
   c. go over first page
   d. write name of contact person and schedule on front
   e. Monday's schedule in detail -
      (TD and SD will not be concise)

Name of Observer Headquarters Phone District

SCHEDULE

Date		Time	Location	Activity
Granite/Nebo	Cache			
3/26	3/26	9:00 - 3:00	Sirloin Stockade, SLC	Training
3/29	4/5	8:30 - 10:00	District Schools	Data collecting
3/29	4/5	1:00 - 3:00	Headquarters	Retraining
3/30	4/6	8:30 - 12:00	District Schools	Data collecting
3/31	4/7	8:30 - 12:00	District Schools	Data collecting
4/01	4/8	8:30 - 12:00	District Schools	Data collecting
4/01	4/8	1:00 - 3:00	Headquarters	Training
4/02	4/9	8:30 - 3:00	District Schools	Data collecting
4/02	4/9	3:00 - 4:00	Headquarters	Final meeting

SUBTESTS FOR OBSERVATION

District	Test	Subtest	Time	Teacher Directed	Timed Test
Granite	SAT	Word Study Skills: Part A	10	X	
		Reading: Part B	25		X
Cache	CTBS	Word Attack	38	X	
		Reading Comprehension	28		X
Nebo	ITBS	Word Analysis	20	X	
		Stories	15		X

NOTES ON OBSERVATIONS

1. Each test (both teacher directed and timed tests) will be observed in each classroom.
2. Observers will be randomly paired each day.
3. During each observation, paired observers will collect data first on the teacher directed, then on the timed tests. Tests will be administered consecutively with a 5-10 minute break between.
4. On Monday, observers will practice for 1 hour in a classroom before retraining that afternoon at headquarters.
5. Data will be collected in the schools on Tuesday, Wednesday, and Thursday mornings.

6. Observers are to return to headquarters each morning after observations have been completed. Forms will be checked and observers will be given new forms and equipment for the next day.

7. On Thursday afternoon, observers will be trained to administer the test-wiseness and student attitude scale. This scale will be administered on Friday.

8. A final meeting is scheduled on Friday afternoon.

9. Checks will be mailed to you on May 10th.

OBSERVATION PROCEDURES

1. Locate the schools before the day you observe. Actually drive to any schools which you may have difficulty finding in a hurry.
2. Fill out forms with the information that you have. Remember to bring the tape recorder, earphone, pencils, tape recording, forms, and clip board to the school.
3. After driving to your assigned school, leave extraneous items (e.g., coats, purses, notebooks, etc.) in the car if possible.
4. Report to office to ask for directions to the teacher's room.
5. Report to the teacher's room and check to see if the subtests are scheduled correctly (first the teacher directed, then the timed test).
6. Arrange your seating so that students can be clearly seen.
7. Set up tape recorder and earphones.
8. Select students to observe from those closest to you. Try to select a representative group by counting off every third student.
9. Identify students on observation form by hair, shirt, dress, etc., and coordinate your observation pattern with your partner.
10. Start to fill in the checklist and keep it handy for notes throughout subtest.
11. Begin taking interval data for the teacher directed test when teacher starts reading the directions from the manual. If the teacher gives students a five-minute break between subtests, do not record data. Begin taking interval data for the timed tests when the teacher starts reading the directions.
12. Remember to break eye contact with students who look at you.
13. Don't show data collection forms to teacher--they are naive to experimental conditions.
14. When both subtests are finished, obtain the names of the observed students from teacher and complete the checklist.
15. Exit from the room as quickly and quietly as possible.
16. Go to next classroom or headquarters to report.

Notes for trainer (attitude and TW)

1. Stand in front of class (make sure you can observe all students)
2. All students must be seated and facing front of class before administering forms.
3. Students need to have sharpened pencil and eraser on top of desk.
4. Students must follow directions (very important)
    a. Allow time for questions.
    b. Make sure students put names on booklets.
5. Pacing of quesitons critical
    a. Allow time for questions
    b. Allow reasonable time for completion of item(s)
6. Discuss class response cue

Order of training

1. Model the administration of each form. Observers should acually work as though they are the students.
2. Discuss the notes above and procedures for administration.
3. Supervise the observers as they practice administrating both forms. (All observers should administer all items.)
4. Distribute envelopes, rubber bands, and extra forms.
5. Schedule debriefing meeting on Friday afternoon.

GENERAL PROCEDURES FOR ADMINISTERING ATTITUDE AND TEST-WISENESS INSTRUMENTS

1. You will administer three forms:

   a. Student attitude
   b. Student test-wiseness
   c. Teacher attitude

2. When referring to forms, call them "booklets".

3. At appropriate breaks in the administration, praise students for working hard, trying their best, listening to instructions, and paying attention.

4. Make sure the teacher's name is on the teacher attitude form.

5. Use group response to obtain answers to questions.

6. Stand in front of the class when giving directions or reading items. Make sure you can see all the students' faces.

7. Clarify and repeat directions (if necessary) and items for student attitude form.

8. Clarify and repeat directions (if necessary) for test-wiseness forms. Do not repeat or explain any items on test-wiseness forms. Tell students to try their best if they want help.

9. Proceed in this order:

   a. Give the teacher attitude form to the teacher before starting with the students.

   b. Introduce yourself to the class with your name and purpose. (For instance, "We want to find out how second grade students feel about tests.")

   c. Have students put a pencil and an eraser on top of their desks.

   d. Pass out and administer student attitude booklet.

   e. Collect student attitude booklet.

   f. Pass out and administer student test-wiseness booklet.

   g. Collect student test-wiseness booklet.

Procedures for Friday

1. You will administer three forms:

    a. Student attitude
    b. Student testwiseness
    c. Teacher attitude.

2. Give the teacher attitude form to the teacher before starting with the students.

3. Introduce yourself with your name and purpose. (For instance, "We want to find out how second grade students feel about tests.")

4. Have students put a pencil and an eraser on top of their desks.

5. Pass out and administer student attitude form.

6. Collect student attitude form.

7. Pass out and administer student testwiseness form.

8. Collect student testwiseness form.

9. When referring to forms, call them "booklets."

10. At appropriate breaks in the administration, praise students for working hard, trying their best, listening to instructions, and paying attention.

SCHEDULE FOR OBSERVERS
GRANITE

Group	School	Teacher	VISIT I							VISIT II			
			3/29	3/30		3/31		4/1		4/2			
			W	T1	T2	W1	W2	T1	T2	F1	F2	F3	F4
E1	Hillsdale 3275 W 3100 S (969-9845)	Jensen 08		9:30 Harris Stephensen						9:30 Janelli			
		Kane 09			11:00 Harris Stephensen					9:30 Fuller			
		Kunz 10		9:30 Foust Rabe						9:30 Akagi			
		Waldram 11			Foust Rabe					9:30 S. Christenson			
	Lincoln [illegible] (266-4485)	Archer 12				9:00 Banks Blackburn					10:15 Turnblue		
		Norris 13					9:55 Banks Blackburn				11:00 S. Christenson		
	West Kearns 490 S 4720 W (968-1113)	Banks [illegible]				9:00 C. Christenson Janelli					11:00 Akagi		
		Borden 15					10:00 C. Christenson Janelli					11:30 Akagi	
		Gomez 16							11:00 Foust Blackburn			11:15 Fuller	
		Green 17						10:00 Foust Blackburn				11:30 Rabe	
		Lobb 18						9:00 Harris Stephensen					1:00 Turnblue
		Martin 19							10:00 Harris Stephensen				1:00 S. Christenson
	Redwood 2650 S. Redwood (972-6065)	Crockett 20	9:30 Stephensen Fuller										1:30 Stephensen
		Latham 21	9:30 Foust Janelli										1:30 Fuller
E2	Western Hills 5190 S. 4620 W. (969-9897)	Cannon 28		9:15 Turnblue S. Christenson									1:00 Janelli
		Eber 29			10:00 Turnblue S. Christenson								1:00 Blackburn
		Tanner 30		9:15 Akagi Fuller									1:00 C. Christenson
		Shepherd 31			10:00 Akagi Fuller							11:30 Harris	
		Schmidt 32						9:15 S. Christenson Akagi				11:30 Banks	
E2	Stansbury 3050 S. 2700 W (972-8932)	Hunt 33				9:00 Fuller S. Christenson					10:00 Rabe		
		Miller 34					10:50 Fuller S. Christenson				10:00 Blackburn		
		Wallace 35				9:00 Akagi Turnblue					10:00 Stephenson		
	South Kearns 4430 W 5570 S (969-9875)	Franco 38					11:00 Akagi Turnblue			9:00 Turnblue			
		Grose 36						9:00 Rabe C. Christenson		9:00 Rabe			
		Madsen 37							10:30 Rabe C. Christenson	9:00 Foust			
C	Lake Ridge 7400 W 3400 S (250-5803)	Belliston 51		9:15 Janelli Blackburn									1:00 Banks
		Woodland 52		10:00 Janelli Blackburn									1:00 Foust
		Spackman 53			11:00 C. Christenson Banks							11:20 C. Christenson	
	Roosevelt 3225 S 800 E (486-0717)	Burton 55				9:00 Rabe Stephensen					10:00 Harris		
		Pugh 54				9:00 Foust Harris					10:00 Banks		
	Wilson 2825 S 200 E (487-1759)	Cummings 57							10:45 Turnblue Fuller		10:00 C. Christenson		
		Jackson 58						9:15 Janelli Banks		9:00 Stephensen			
		Lund 56							10:45 Janelli Banks	9:00 Harris			
	Hillview 1035 E 4500 S (261-1417)	Ginger Rhode	9:45 Rabe Blackburn Turnblue C. Christenson										
	Morningside 4170 S 3000 E (272-2012) Room 10	Geneva Shanks	10:45 S. Christenson Banks- Harris Akagi										
SUBSTITUTE				C. Christenson Banks	Janelli Blackburn		Rabe Stephensen	Turnblue Fuller	Akagi S. Christenson	Blackburn Banks	Fuller Janelli	Turnblue S. Christenson	Akagi Harris

BEST COPY AVAILABLE

SCHEDULE FOR OBSERVERS
CACHE

			PRACTICE	VISIT I						VISIT II			
			3/29	4/6		4/7		4/8		4/9			
roup	School	Teacher	M	T1	T2	W1	W2	T1	T2	F1	F2	F3	F4
E1	Wellsville 90 E 100 S (245-3764)	Jenkins 01				8:30 Turner Godfrey					9:30 Godfrey		
		Murray 03				8:30 Garner Jenkins							1:30 Jenkins
		Nielsen 02					9:45 Garner Jenkins			8:30 Godfrey			
E2	Lewiston 107 E 200 S (258-2923)	Mieure 22						8:30 Meikle Godfrey			9:30 Meikle		
		Schenevar 23							10:00 Meikle Godfrey	8:30 Meikle			
C	Millville 67 S. Main (752-7162)	Noble 40		8:15 Godfrey Meikle								10:15 Turner	
		Tuddenham 39			9:45 Garner Turner								1:30 Turner
	Summit 80 W. Center Smithfield (563-6269)	Jensen 41		10:00 C Godfrey Meikle				10:00 WA Turner Jenkins		8:30 Garner			
		Mellville 43					Comp&Math		8:30 Turner Jenkins		9:30 Garner		
		Rawlins 42					10:15 Turner Godfrey					10:30 Jenkins	
	Hillcrest Logan 752-3941	Peterson	10:00 Jenkins Godfrey										
		Olsen	10:00 Garner Turner										
		Larsen	10:00 Jenkins Taylor										
	SUBSTITUTE			Jenkins	Jenkins	Meikle	Meikle	Garner	Garner	Turner	Turner	Godfrey	Godfrey

SCHEDULE FOR OBSERVERS
NEBO

			PRACTICE	VISIT I						VISIT II			
			3/29	3/30		3/31		4/1		4/2			
roup	School	Teacher	M	T1	T2	W1	W2	T1	T2	F1	F2	F3	F4
E1	Westside 500 S. Main Springville (489-6101)	Burbidge 04		9:00 Burnham Howard								10:45 Scovill	
		Payne 05			10:30 Burnham Howard								12:45 Howard
	Santaquin 74 W 100 S (754-3611)	Anthony 07						8:45 Stewart Reed			9:45 Burnham		
		Willis 06							10:10 Stewart Reed	8:45 Burnham			
E2	Wilson 590 W 500 S Payson (465-3182)	Altenburg 25				8:45 Howard Reed					9:45 Stewart		
		Anderson 24					10:15 Howard Reed			8:45 Stewart			
	Goshen 10 N Center (667-3361)	Boyack 27						8:30 Scovill Burnham					12:30 Reed
		Neff 26							10:00 Scovill Burnham			10:45 Burnham	
C	Larsen 1175 E. Flonette Spanish Fork (798-9520)	Jensen 44		8:30 Reed Scovill							9:15 Howard		
		Lee 45		9:30 Reed Scovill						8:30 Scovill			
		Smith 46			10:30 Reed Scovill	9:15 JC Burton				8:30 Howard			
	Taylor 40 S 500 W Payson (465-2231)	Beaudin 47				8:45 Scovill Stewart							12:30 Stewart
		Ghiradelli 48					10:00 Scovill Stewart						12:30 Burnham
	Brookside 750 E 400 S Springville (489-4241)	Mason 49	9:00 Burnham Scovill Reed									10:30 Howard	
		Lee 50	9:00 Howard Stewart								9:30 Reed		
	SUBSTITUTE			Stewart	Stewart	Burton	Burton	Howard	Howard	Reed	Scovill	Reed	Scovill

Instructions: Please read each question carefully and circle the response which best indicates your feelings about standardized achievement tests. There is no need to write your name on this questionnaire, because only group responses will be analyzed. There are no right or wrong answers, so please be as honest and candid as possible. Thank you for your help.

Means and SD on Each Item and Subscale for Each Group				Percentage of Respondents Selecting Each Option							
EI	EII	C	Total								
				I.	**What is your general opinion of standardized achievement tests?**						
3.00 .95	3.00 .75	3.10 .91	3.05 .87	1.	Harmful for students	5.2 1	13.8 2	56.9 3	19.0 4	5.2 5	Helpful for students
3.24 1.09	3.41 1.06	3.10 1.16	3.24 1.10	2.	Not useful for teachers	6.9 1	19.0 2	27.6 3	36.2 4	10.3 5	Very useful for teachers
2.71 1.23	2.71 1.05	2.90 .79	2.78 1.03	3.	Unfair	12.1 1	24.1 2	43.1 3	15.5 4	5.2 5	Fair
2.52 1.12	2.65 .79	2.75 .85	2.64 .93	4.	Invalid	13.8 1	24.1 2	48.3 3	12.1 4	1.7 5	Valid
3.29 .96	3.29 .92	3.05 1.05	3.21 .97	5.	Too difficult for students	1.7 1	20.7 2	44.8 3	20.7 4	12.1 5	Appropriately difficult for students
2.94 .88	3.02 .72	2.98 .62	2.98 .74		Total (1-5)						
				II.	**How do you feel about administering standardized achievement tests?**						
3.71 1.01	4.18 1.02	3.50 1.19	3.78 1.09	6.	Anxious	1.7 1	13.8 2	20.7 3	32.8 4	31.0 5	Calm
4.05 .74	4.00 .94	4.05 .76	4.04 .79	7.	Uninterested	0.0 1	0.0 2	29.3 3	37.9 4	32.8 5	Interested
4.52 .60	4.29 .69	4.05 .83	4.29 .73	8.	Not knowledgeable	0.0 1	1.7 2	10.3 3	44.8 4	43.1 5	Knowledgeable
3.95 .81	3.94 .90	3.80 .95	3.90 .87	9.	Antagonistic	0.0 1	5.2 2	27.6 3	39.7 4	27.6 5	Supportive
4.43 .68	4.59 .51	3.95 .83	4.31 .73	10.	Insecure	0.0 1	1.7 2	10.3 3	43.1 4	44.8 5	Confident
4.12 .54	4.20 .58	3.86 .62	4.06 .58		Total (6-10)						

II. Standardized achievement test results are used in many ways by different teachers and school systems. Please indicate how useful you think such test results could be for each of the following purposes:

Means and SD on Each Item and Subscale for Each Group					Percentage of Respondents			
EI	EII	C	Total		Harmful	Not Useful	Somewhat Useful	Very Useful
2.86 .79	2.82 .73	2.90 .97	2.90 .83	11. To report to parents to help them interpret their child's performance in school.	8.6 1	15.5 2	56.9 3	19.0 4
3.05 .67	3.12 .33	3.00 .80	3.05 .63	12. To measure the educational status of individual students as compared with others their age.	1.7 1	12.1 2	65.5 3	20.7 4
2.95 .59	2.77 .66	3.05 .51	2.93 .59	13. To measure the educational "growth" of students from year to year.	1.7 1	15.5 2	70.7 3	12.1 4
2.71 .72	2.88 .70	2.90 .64	2.83 .68	14. To screen students for and make decisions about placement in special education programs.	3.4 1	22.4 2	62.1 3	12.1 4
2.67 .66	2.88 .70	2.85 .81	2.79 .72	15. To help plan instruction for individual students.	1.7 1	32.8 2	50.0 3	15.5 4
2.86 .73	2.88 .70	3.05 .76	2.93 .72	16. To help plan instruction for class groups.	1.7 1	24.1 2	53.4 3	20.7 4
2.48 .68	2.65 .70	2.55 .51	2.55 .63	17. To evaluate specific teaching methods, instructional materials, and/or educational programs.	3.4 1	41.4 2	51.7 3	3.4 4
1.71 .90	1.82 .73	1.60 .75	1.71 .80	18. To report to newspapers informing the public about differences between schools.	48.3 1	34.5 2	15.5 3	1.7 4
2.10 .83	2.35 .79	2.45 .76	2.29 .80	19. To report to administrators as an aid in decision making.	19.0 1	34.5 2	44.8 3	1.7 4
1.57 .75	1.47 .72	1.50 .61	1.52 .68	20. To evaluate and make comparisons between the performance of different teachers.	58.6 1	31.0 2	10.3 3	0.0 4
2.49 .49	2.42 .68	2.58 .42	2.54 .43	Total (11-20)				

EI	EII	C	Total	IV. Would you personally be in favor of:	Strongly in favor of	Somewhat in favor of	Somewhat against	Strongly against
1.95 .81	2.12 .78	2.10 .85	2.05 .80	21. Increased use of "minimum competency tests" to determine high school graduation?	22.4 1	56.9 2	13.8 3	6.9 4
3.10 .94	3.12 .70	3.05 .61	3.09 .76	22. Increased use of achievement test results to compare how successful various schools are?	1.7 1	19.0 2	48.3 3	31.0 4
2.29 .72	2.41 .71	2.70 .98	2.47 .82	23. Increased use of achievement test results for feedback to students about their performance?	6.9 1	53.4 2	25.9 3	13.9 4
2.05 .81	2.29 .59	2.00 .80	2.10 .74	24. Increased use of achievement test results by classroom teachers to make instructional and curriculum decisions?	20.7 1	50.0 2	27.6 3	1.7 4
2.65 .65	2.40 .55	2.52 .62	2.57 .59	Total (21-24)				

BEST COPY AVAILABLE

Means and SD on Each Item
and Subscale for Each Group

Percentage of Respondents

V. How do you think standardized achievement tests make your students feel?

								EI	EII	C	Total
25.	Anxious	15.5 1	46.6 2	25.9 3	12.1 4	0.0 5	Calm	2.62 .97	2.18 .88	2.20 .77	2.35 .89
26.	Smart	1.7 1	19.0 2	41.4 3	32.8 4	5.2 5	Dumb	3.14 .85	3.00 .87	3.45 .89	3.21 .87
27.	Bad	5.2 1	17.2 2	56.9 3	17.2 4	3.4 5	Good	3.29 .96	2.94 .61	2.65 .75	2.97 .84
28.	Afraid	8.6 1	34.5 2	37.9 3	10.3 4	8.6 5	Not afraid	3.10 1.22	2.65 .86	2.50 .95	2.76 1.05
29.	Successful	5.2 1	17.2 2	50.0 3	20.7 4	6.9 5	Unsuccessful	2.81 1.12	3.18 .81	3.25 .79	3.07 .93
30.	Insecure	5.2 1	22.4 2	53.4 3	15.5 4	3.4 5	Confident	3.29 .90	2.76 .75	2.60 .75	2.90 .85
Total (25-30)								3.05 .83	2.71 .65	2.53 .56	2.78 .72

VI. This is the end of the questionnaire concerning teachers' attitudes towards standardized tests. To assist us in analyzing the other data from the project, we would like you to answer a few more questions concerning standardized test administration procedures in your classroom. Again, there are no right or wrong answers.

Part VI responses are summarized on Quality of Test Administration Checkli

1. When did you first tell your students they would be taking a standardized achievement test?

   a. the day the test began
   b. the day before the test began
   c. 2-5 days before the test began
   d. other, please specify ______________________________

2. Did you do anything in particular to prepare the students for taking the standardized test?

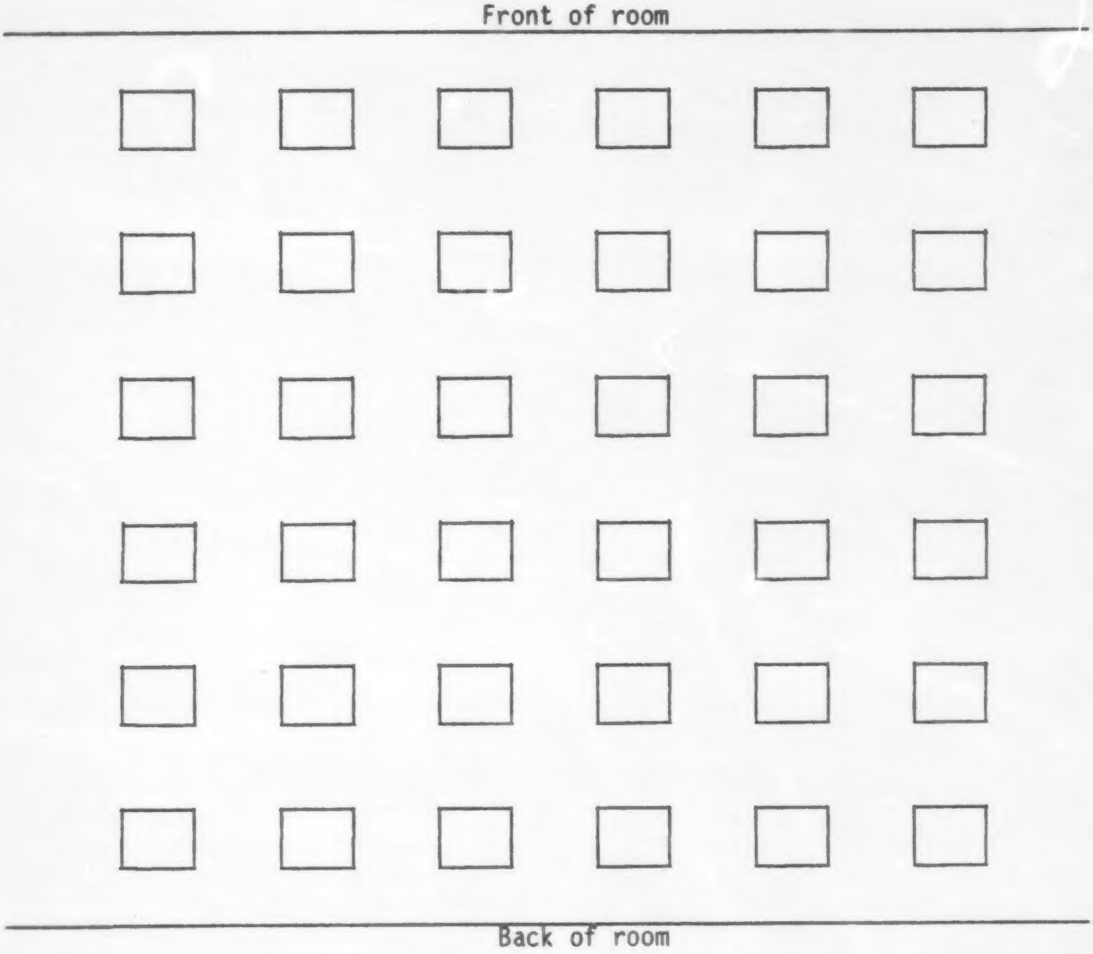
   a. No b. Yes (please explain) ____________________

3. During the standardized test, did you give students any specific instructions about what they should do if they finished a timed subtest before the allotted time was up?

   a. No b. Yes (please explain) ____________________

BEST COPY AVAILABLE

4. The diagram below is how a typical classroom might be arranged. For the 5 children in your class who have the most difficulty with misbehaving, acting out, or lack of attention, indicate with "X's" (one for each child) the approximate location of where they sat during the standardized test. (Note: even though your room may not have individual desks or the desks may be arranged differently, you can still approximate the location of these students using this diagram.



STUDENT ATTITUDE TOWARDS STANDARDIZED TESTING
DIRECTIONS

DO	SAY
1. Print this on the blackboard. fun 0——0——0 boring good 0——0——0 bad	
2. Demonstrate where to put pencil.	I will pass a booklet to each of you. Put your pencils on the booklet. Do not turn the booklet over until I tell you to.
3. Pass out copies of booklets faced down to each student.	
4.	Today I will ask you some questions about how you feel toward the test you have been taking this week. What we will do today is not really a test because there are no right or wrong answers. I will read part of a sentence, then you will mark the answer to finish the sentence.
5.	First, I will show you how to mark your answers. Listen to this sentence. I think playing baseball is. Say that with me. I think playing baseball is.
6. Point to the first line on the board.	Look at the board. Here are some answers and circles. At one end is a circle near the word "fun." At the other end is a circle near the word "boring." The circle in the middle means that playing baseball is not fun and not boring but sort of in between.

DO	SAY
7. Point to each circle. Fill in circles with side-to-side motion.	Who thinks that playing baseball is fun? Then you would have marked this circle. How many think playing baseball is in between fun and boring? Then you would have marked this circle. How many think playing baseball is boring? Then you would have marked this circle. Remember, fill in only one circle for each sentence.
8.	Some people think baseball is fun and others think it is boring or in between. There is no right or wrong answer for this question. Is there a right or wrong answer to this question? (Students: No.)
9. Demonstrate. Pause and check fingers.	Turn your paper over. (Pause) Look at the first page. Put your finger on page number 1 at the bottom.
10. Check names.	Put your finger on the word "name" at the bottom. Write your first and last name on the line.
11. Pause and check fingers.	Now, put your finger on sample number 1 at the top of the page. A sample shows you how to do other items. What does a sample show you? (Students: How to do the other items.)
12. Check fingers. Pause and wait for students to mark paper.	For sample number 1, I will read a sentence and you will mark the circle that tells best how you feel. The sentence is, I think broccoli tastes. Say that with me. I think broccoli tastes. Now, mark the circle that tells you how you think broccoli tastes.
13. Demonstrate on board. Fill in circles.	If you think broccoli tastes good, you should have marked this circle. If you think broccoli tastes bad, you should have marked this circle. And, if you think broccoli tastes kind of in between good and bad, you should have marked this circle.

	DO	SAY
14.		Different people think broccoli tastes different, so there is no right or wrong answer. Is there a right answer for this sentence? (Students: No.) That's right, there is no right or wrong answer for this sentence.
15.	Check fingers. Pause and wait for students to mark paper.	Put your finger on sample number 2. Listen. The sentence is, Math problems are. Say it with me. Math problems are. Math problems are good, in between, bad. Now, mark the circle that tells best how you feel about math problems.
16.	Check fingers.	Now, turn the page to page number 2 and put your finger on page number 2 at the bottom so I can see you are on the right page. Good. For these items, you will mark the circle that tells how you feel about the test you have been taking this week. Remember, there are no right or wrong answers. Are there any right or wrong answers to these questions? (Students: No.)
17.		For each item, you should mark a circle that is near the word that tells how you feel. Who knows which circle you should mark? (Student: Near the word that tells how I feel.) Raise your hand and ask me if you have any questions.
18.	Repeat any items that seem confusing and explain words that students do not understand.	Good. Point to item number 1. Taking tests makes me feel not afraid, in between, afraid. Mark the circle that shows best how taking tests makes you feel.
19.		Finger on item number 2. Taking tests makes me feel happy, in between, sad. Mark the circle that shows best how you feel.
20.		Item number 3. Taking tests makes me feel smart, in between, dumb.

DO	SAY
21.	Item number 4. Taking tests makes me feel good, in between, bad.
22.	Item number 5. Taking tests makes me feel calm, in between, nervous.
23.	Item number 6. I think tests are fun, in between, boring.
24.	Item number 7. I think tests are fair, in between, not fair.
25.	Item number 8. Do tests help your teacher teach you better? Yes, in between, no.
26.	Check your paper to see that you have marked a circle for every item. Now, pass your papers to the front of the room and I will collect them.

What Do You Think ?

Sample # 1

good ○———○———○ bad

Sample # 2

easy ○———○———○ hard

Name ______________________________

1	not afraid	○	○	○	afraid
2	happy	○	○	○	sad
3	smart	○	○	○	dumb
4	good	○	○	○	bad
5	calm	○	○	○	nervous
6	fun	○	○	○	boring
7	fair	○	○	○	not fair
8	yes	○	○	○	no

MEAN SCORE
STANDARD DEVIATION

EI	EII	C	Total
1.3	1.4	1.4	1.4
.6	.6	.6	.6
p = .04			
1.4	1.4	1.5	1.4
.6	.5	.6	.6
p = .08			
1.3	1.3	1.3	1.3
.6	.6	.6	.6
1.4	1.4	1.5	1.4
.7	.6	.7	.6
p = .05			
1.9	2.0	2.0	2.0
.9	.9	.9	.9
1.9	1.7	2.0	1.9
.9	.8	.9	.9
p = .0			
1.4	1.3	1.4	1.4
.7	.6	.6	.7
1.3	1.2	1.4	1.3
.6	.5	.7	.6
p = .0			

Percent of Students
Selecting Each Option

1	not afraid ○	66.6 ○	28.7 ○	4.6 ○	afraid
2	happy ○	61.2 ○	32.7 ○	6.0 ○	sad
3	smart ○	77.4 ○	15.0 ○	7.6 ○	dumb
4	good ○	67.0 ○	24.6 ○	8.4 ○	bad
5	calm ○	39.7 ○	20.8 ○	39.4 ○	nervous
6	fun ○	48.6 ○	17.3 ○	34.1 ○	boring
7	fair ○	73.7 ○	16.4 ○	9.7 ○	not fair
8	yes ○	81.5 ○	9.2 ○	9.0 ○	no

Do tests help your teacher teach you better?

E1	E2	C	Total	
				6. Which of the following animals does *not* lay eggs?
6.7	6.7	4.1	5.8	○ chickens
31.2	27.5	31.5	30.3	○ snakes
4.7	2.8	4.5	4.1	○ birds
56.9	61.9	59.2	59.2	● orangutans
0.4	1.0	0.6	0.7	Blank
				7. Abraham Lincoln, who lived during the Civil War, was
67.1	65.5	68.0	67.0	● a president of the United States
11.0	8.8	9.7	9.9	○ a multimillionaire
7.3	6.5	6.4	6.8	○ a pilot
13.9	16.8	14.6	15.0	○ a soldier
0.6	2.3	1.3	1.3	Blank
				8. Which of the following animals does lay eggs?
26.3	20.5	24.2	23.9	○ gnus
50.2	50.3	49.8	50.1	● snakes
11.2	13.5	12.0	12.1	○ sheep
11.8	14.0	13.3	13.0	○ mice
0.4	1.8	0.6	0.9	Blank
				9. What is a hyperbola?
26.9	28.5	33.3	29.6	○ a planet
29.8	26.9	27.3	28.1	● the locus of a point whose difference in distances from two fixed points is constant
17.6	16.8	13.9	16.1	○ burned wood
23.3	23.8	23.4	23.5	○ a satellite
2.4	3.9	2.1	2.8	Blank
				10. Dwight Eisenhower
45.7	43.0	45.3	44.8	● was a general during World War II
14.1	16.1	20.2	16.8	○ astronaut
19.8	17.4	15.5	17.6	○ of Russia
17.8	19.7	16.5	17.9	○ in the United States
2.7	3.9	2.6	3.0	Blank
				11. Sally and Jane are in the same class. Sally can hang from the bar for 10 seconds. Jane can hang from the bar a lot longer than anyone in her class. She hangs for
7.8	8.3	6.7	7.5	○ 2 seconds
22.9	22.0	20.4	21.8	○ 10 seconds
15.1	12.7	15.2	14.5	○ 30 seconds
53.1	54.9	56.0	54.6	● 60 seconds
1.2	2.1	1.7	1.6	
				12. "Four score and seven years ago . . ." is the beginning of the
22.4	22.5	19.7	27.0	○ Papa Encyclical Rerum Novarem
28.4	26.2	30.3	25.9	○ President Reagan's State of the Union message
18.2	18.4	19.5	18.4	○ The Gettysburg Address
28.2	28.8	27.0	26.0	○ John Kennedy's funeral eulogy
2.9	4.1	3.4	2.8	Blank

GO ON

CLASS EXERCISE

PART A

Sample: On a nice day, the sky is

○ white
● blue
○ pink
○ black

E1	E2	C	Total	
				1. An anesthesiologist is a
26.1	23.3	26.6	25.5	○ basketball player
15.1	15.8	16.5	15.8	○ barber
43.7	42.0	45.3	43.7	● physician
13.9	17.6	9.9	13.6	○ hairdresser
1.2	1.3	1.7	1.4	Blank
				2. What sport is played on a field?
34.9	46.4	36.3	38.7	○ soccer
47.6	40.9	50.2	46.6	○ football
4.1	4.4	3.0	3.8	○ polo
13.3	8.0	10.3	10.7	● all of the above
0.2	0.3	0.2	0.2	Blank
				3. Who discovered how to pasteurize milk?
19.8	20.7	20.4	20.3	○ Madame Pompadour
16.9	18.1	17.2	17.4	○ Madame Curie
31.2	26.9	33.3	30.7	● Louis Pasteur
31.0	32.6	28.1	30.5	○ Milo Bishop
1.0	1.6	1.1	1.2	Blank
				4. Herringbone is
33.3	29.0	35.4	32.8	○ a fossil
22.2	19.4	23.0	88.7	● pickled mackerel that has been frozen
21.4	20.2	18.7	20.1	○ a fabric pattern
21.6	28.8	21.5	23.6	○ a small animal that lives on Mars
1.4	2.6	1.5	1.8	Blank
				5. About how many glasses of water should a person drink each day?
26.1	23.8	24.0	24.7	○ 1
6.9	7.5	7.9	7.5	○ 30
52.2	59.1	58.6	56.4	● 8
14.1	8.5	9.0	10.7	○ 50
0.6	1.0	0.4	0.7	Blank

GO ON

E1	E2	C	Total	
				13. The number of men who have been president of the United States is less than
58.4	62.2	60.3	60.1	● 45
14.3	14.8	14.2	14.4	○ 30
11.0	8.5	7.9	9.2	○ 18
13.7	13.0	13.9	13.6	○ 7
2.7	1.6	3.6	2.7	Blank
				14. A stretch of land between two mountains is called a:
20.4	20.2	16.7	19.1	○ hill
15.1	12.2	11.2	12.9	○ river
14.9	13.0	12.4	13.5	○ mound
46.1	52.1	56.0	51.3	● valley
3.5	2.6	3.6	3.3	Blank
				15. The average person lives about
18.6	13.7	14.4	15.7	○ 20 years
31.2	39.6	40.1	36.7	● 73 years
22.0	17.9	21.0	20.5	○ 150 years
24.1	26.4	20.2	23.4	○ 200 years
4.1	2.3	4.3	3.7	
				16. Abalone is an ocean crustacean. It lives in the
13.7	10.6	12.7	12.4	○ trees
16.1	17.4	15.9	16.4	○ ground
18.0	20.7	17.0	18.4	● lake
46.9	47.4	49.6	48.0	○ sea
5.3	3.9	4.9	4.8	Blank
				17. The number of miles from the earth to the moon is less than
18.8	14.0	15.7	16.3	○ 243,000
43.7	50.5	45.7	46.3	● 250,000
8.6	9.8	6.7	8.3	○ 244,000
24.3	22.5	26.8	24.7	○ 249,000
4.7	3.1	5.2	4.4	Blank
				18. Abalone is
23.3	19.9	25.5	23.1	○ a disease of one foot
35.9	36.5	36.9	36.4	● a crustacean
20.2	22.5	18.7	20.3	○ a cold cut
14.1	15.0	12.4	13.8	○ a style of hair
6.5	6.0	6.4	6.3	Blank
				19. The Susan B. Anthony dollar honors
24.3	19.9	22.3	22.4	○ one of our founding fathers
31.2	31.9	36.3	33.2	● the woman who led the suffragette movement
17.8	18.1	17.6	17.8	○ a famous baseball player
20.0	23.8	16.3	19.8	○ the husband of Betsy Ross
6.7	6.2	7.5	6.9	Blank

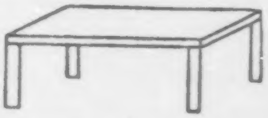
GO ON

E1	E2	C	Total	
				20. The equator
48.6	43.0	42.5	44.9	● is hot all year round
17.1	14.5	15.5	15.8	○ mossy rocks
15.3	22.8	19.1	18.8	○ pieces of lava rock
11.0	12.4	12.2	11.8	○ thick black smoke
8.0	7.3	10.7	8.7	Blank
				21. The Star Spangled Banner, written by Francis Scott Key, is
36.7	44.0	43.3	41.1	● the national anthem
18.8	16.3	20.0	18.5	○ a poem
22.9	19.4	14.2	18.9	○ a book
12.4	13.0	12.0	12.4	○ a magazine
9.2	7.3	10.5	9.1	Blank
				22. Which is not a flower?
11.8	9.6	6.0	9.2	○ rose
26.3	22.5	26.8	25.4	○ paisley
10.8	9.3	10.3	10.2	○ tulip
41.2	52.3	44.6	45.6	● goldenrod
9.8	6.2	12.2	9.6	Blank
				23. What does erosion do?
17.1	18.9	18.2	18.0	○ carries light
26.9	23.8	28.1	26.5	○ makes oxygen
14.3	17.9	12.4	14.7	○ gives flowers water
29.4	29.5	27.7	28.8	● carries away the material that is broken up by weathering
12.2	9.8	13.5	12.0	Blank
				24. Which of the following are vegetables?
23.5	24.9	25.5	24.6	○ corn
13.9	13.5	13.1	13.5	○ rutabaga
18.2	27.2	21.0	21.8	○ carrots
32.2	23.3	25.8	27.4	● all of the above
12.2	11.1	14.6	12.7	Blank
				25. The capital of Oklahoma is
12.4	11.1	10.7	11.5	○ Tulsa
45.9	50.8	51.3	49.2	● Oklahoma City
13.1	13.7	9.2	11.9	○ Phoenix
14.5	12.2	11.6	12.8	○ Salem
14.1	12.2	17.2	14.6	Blank
				26. Tell which movie was not about the future.
18.2	19.2	13.9	17.0	○ 2001 Space Odyssey
27.6	24.6	23.4	25.3	● 1776
15.9	17.6	17.2	16.8	○ Star Trek
25.5	25.1	28.5	26.5	○ Star Wars
12.9	13.5	17.0	14.5	Blank

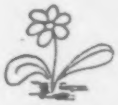
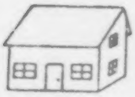
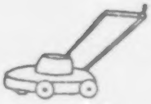
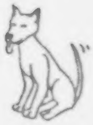
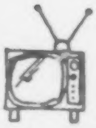
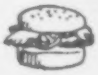
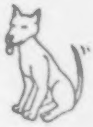
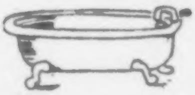
STOP

PART B

Sample:



0 top
0 leg
0 table
0 wood

E1	E2	C	Total			
12.4	10.1	15.7	12.9	1.		0 plant
0.2	0.3	0.2	0.2			0 boy
0.0	0.0	0.0	0.0			0 house
86.3	88.3	83.7	86.0			● flower
1.0	1.3	0.4	0.9			Blank
15.5	15.5	17.0	16.0	2.		0 paper
27.1	30.3	27.9	28.3			0 sidewalk
53.3	51.3	50.9	51.9			● abode
2.4	1.8	2.1	2.2			0 ball
1.6	1.0	2.1	1.6			Blank
2.2	1.3	2.8	2.2	3.		0 cut
93.7	95.3	93.6	94.1			● lawn mower
1.4	0.8	0.9	1.0			0 blade
2.4	1.8	2.6	2.3			0 grass
0.2	0.8	0.2	0.4			Blank
6.5	6.0	9.2	7.3	4.		0 animal
0.2	0.3	1.1	0.5			0 paw
0.4	0.5	0.4	0.4			0 hair
97.2	92.5	89.3	91.4			● dog
0.2	0.8		0.3			Blank
3.9	4.4	4.9	4.4	5.		0 watch
1.6	1.3	1.5	1.5			0 look
93.7	93.0	92.9	93.2			● television
0.8	1.3	0.6	0.9			Blank
4.7	3.6	6.2	4.9	6.		0 food
1.0	1.3	2.8	1.7			0 meat
93.7	93.8	90.6	92.6			● hamburger
0.0	0.0	0.0	0.0			Blank
0.6	1.3	0.4	0.7			
72.7	81.1	77.7	76.8	7.		● canine
6.3	5.7	6.4	6.2			0 lamp
5.5	4.1	4.7	4.8			0 chair
12.2	6.2	9.0	9.4			0 house
3.3	2.8	2.1	2.8			Blank
2.2	0.3	0.9	1.2	8.		0 clean
93.5	96.4	95.5	95.0			● bathtub
3.1	1.8	2.1	2.4			0 water
0.4	0.5	1.3	0.7			0 faucet
0.8	1.0	0.2	0.7			Blank

STOP

PART C

The boy, whose name was John, went to the store.

E1	E2	C	Total		
2.4	3.4	2.4	2.7	1.	store 0
4.3	2.6	2.1	3.1		was 0
91.8	92.5	92.7	92.3		went ●
1.0	1.0	1.9	1.3		the 0
0.4	0.5	0.9	0.6		Blank
87.6	86.8	86.5	87.0	2.	the ●
1.5	1.8	1.5	1.6		went 0
1.4	2.8	1.7	1.9		boy 0
9.2	8.0	9.9	9.1		store 0
0.2	0.5	0.4	0.4		Blank
3.9	0.8	2.1	2.4	3.	boy 0
94.3	94.8	93.1	94.0		John ●
1.0	1.8	2.6	1.8		went 0
0.6	2.1	0.6	1.0		to 0
0.2	0.5	1.5	0.7		Blank

E1	E2	C	Total			
1.2	0.3	0.6	0.7	4.	the	0
2.7	2.1	1.5	2.1		name	0
2.0	0.5	0.4	1.0		was	0
93.3	96.4	96.1	95.2		store	●
0.8	0.8	1.3	1.0		Blank	
77.3	81.9	85.8	81.6	5.	went	●
12.4	11.9	8.4	10.9		John	0
7.3	4.1	3.6	5.1		the	0
1.8	1.3	1.3	1.5		store	0
1.0	0.8	0.9	0.9		Blank	
13.5	15.0	16.5	15.0	6.	John	0
67.6	69.4	66.3	67.7		was	●
17.1	13.2	14.2	15.0		boy	0
1.2	1.3	1.5	1.3		to	0
0.6	1.0	1.5	1.0		Blank	

STOP

TEST-WISENESS TEST (CLASS EXERCISE)

DIRECTIONS

DO	SAY
1. Pass out copies of booklets faced down to each student.	Today, we are going to do an exercise together. It is a little like a game. Your job is to try your best to find the right answers to the questions on these sheets. It will be hard because you will not know the right answers to most of these questions. You're not supposed to know all of the answers, but you must try your very best to figure the answers out. Class, will you know the answers to most of these items? (Students: No.) Right. Is your job to try your best to figure out the answers anyway? (Students: Yes.) Good! Every item has only one correct answer. How many correct answers does each item have? (Students: One.) That's right! Let's begin.
	Turn your booklet over.
2. Demonstrate.	
Pause and check fingers.	Put your finger on the number 1 at the bottom of the page. Good!
3.	Now, put your finger on the sample at the top of the page. SAMPLE: On a nice day, the sky is 0 white 0 pink 0 blue 0 black The sample will show you how to do the other items. What does the sample show you? (Students: How to do the other items.) Yes. Read the sample to yourself. (Pause) Now, read the sample sentence with me. On a nice day, the sky is. Good. Now, let's read the four answer choices together. White, blue, pink, black.

	DO	SAY
4.	Check to make sure students marked answer correctly.	Which word is the best answer to the sample question? (Students respond.) Yes, "blue." "On a nice day, the sky is blue." Everyone, mark the space in front of "blue" to show that it is the best answer. (Pause) For each item in Part A, try to find the one best answer and mark it the same way.
5.	Turn to page 4 of test and point to STOP	When I say begin, continue working until you see the word STOP on page 4. You should work until you see what word? (Students: The word "stop.") I will tell you to stop in 10 minutes, so you will have to work very fast.
6.	Check fingers.	Finger on item number 1.
	Time for 10 minutes.	Ready, begin.
	When only 1 minute is left.	You have only one minute left to finish.
7.		Stop. Pencils down.
8.	Demonstrate.	Turn to page 5. (Pause) Put your finger at the top of the page where it says Part B. Now, put your finger on the sample. Good! For this sample, you must find the word that best tells about the picture. Class, what is the picture? (Students: A table.) That's right, a table. Now, read the four answer choices with me. (Students: Top, leg, table, wood.) SAMPLE: Picture of table 0 top 0 leg 0 table 0 wood

DO	SAY
4. Check to make sure students marked answer correctly.	Which word is the best answer to the sample question? (Students respond.) Yes, "blue." "On a nice day, the sky is blue." Everyone, mark the space in front of "blue" to show that it is the best answer. (Pause) For each item in Part A, try to find the one best answer and mark it the same way.
5. Turn to page 4 of test and point to STOP	When I say begin, continue working until you see the word STOP on page 4. You should work until you see what word? (Students: The word "stop.") I will tell you to stop in 10 minutes, so you will have to work very fast.
6. Check fingers.	Finger on item number 1.
Time for 10 minutes.	Ready, begin.
When only 1 minute is left.	You have only one minute left to finish.
7.	Stop. Pencils down.
8. Demonstrate.	Turn to page 5. (Pause) Put your finger at the top of the page where it says Part B. Now, put your finger on the sample. Good! For this sample, you must find the word that best tells about the picture. Class, what is the picture? (Students: A table.) That's right, a table. Now, read the four answer choices with me. (Students: Top, leg, table, wood.) SAMPLE: Picture of table 0 top 0 leg 0 table 0 wood

	DO	SAY
9.	Check to make sure the students fill out the answer space correctly.	Look at the picture again. Which word best tells about the picture? (Students: Table.) Yes, table is the best answer. Put a mark in the space in front of the word "table" to show that it is the correct answer.
10.		Good. When I tell you to begin, you will do all of the items in Part B the same way.
11.	Demonstrate. Time: 2 minutes.	Put your finger on the stop sign at the bottom. Continue working until you come to this stop sign. Finger on item number 1. Ready, begin. Stop. Pencils down.
12.	Check to make sure students are on correct page.	Nice work! Now, for the next part of our exercise. Turn to page 6. (Pause) Look at the top of the page where it says Part C. Are you all at the right place? (Students: Yes.) Great! To do Part C, you will read a sentence and then find the answers to the questions I will ask you. You may look back at the sentence to help you find the right answer. Class, to find the answers to these items, can you look back at the sentence? (Students: Yes.) I will only be able to tell you the question once, so listen carefully.
13.	Check to see if students' fingers are on correct items.	Now, put your finger on the sentence at the top of the page. Read it to yourselves. (Pause) Now let's read it together. ("The boy, whose name was John, went to the store.") Good. Now, put your finger on item 1. Listen to the directions. Mark the word that comes after the word John in the sentence. (Pause) Finger on item number 2. Mark the word that comes before the last word in the sentence. (Pause)

DO	SAY
	Finger on item number 3. Mark the word that begins with a capital letter. (Pause) Finger on item number 4. Mark the word that is the last word in the sentence. (Pause) Finger on item number 5. Mark the word that comes after the second comma. (Pause) Finger on item number 6. Mark the word that comes before the word with the capital letter. (Pause)
14. Check for names.	Good. Pencils down. Close your booklets. Now, class, please write your first and last names on the top of page 1. Good. Thank you for your good work today.

Appendix H

Supplementary Data About Effect of Intervention on Dependent Variables

1. Data File Code Book and Listing of Data
2. Sample Size, Means, Standard Deviations, and Medians for All Dependent Variables Broken Down by Experimental Group for Various Subsamples of Students (H1-H6)
3. Frequencies of Major Variables

MASTER FILE CODE

	Columns	Variable	Label	Code Name
CARD I	1	V01	GROUP C1	Experimental Group
	2-4	V02	ID C2	Student I.D. #
	5-14	-	-	Blank
	15	V03	DIST	District
	16-17	V04	SCH C3	School
	18-19	V05	TCHR C4	Teacher (classroom)
	20	-	-	Blank
	21-22	V06	TATOP C5	Teacher Attitude Opinion of Tests
	23-24	V07	TATFEE C6	Teacher Attitude Teacher Feelings
	25-26	V08	TATUSE C7	Teacher Attitude Use of Tests
	27-28	V09	TATINC C8	Teacher Attitude Increased Use of Tests
	29-30	V10	TATFEEL C9	Teacher Attitude Students Feelings
	31-33	V11	TOTTD C10	Teacher On-Task, Teacher Directed Test
	34-36	V12	TOTSD C11	Teacher On-Task, Student Directed Test
	37-39	V13		Student On-Task, Teacher Directed On-Task
	40-42	V14	SOTTDONP	" " , On-Task Probably
	43-45	V15		Stud. On-Task, Stud. Directed, Tchr. Stop, On-Task
	46-48	V16		Stud. On-Task, Stud. Directed, Tchr. Stop, On-Task Probably
	49-51	V17		Stud. On-Task, Stud. Directed, Stud. Stop, On-Task
	52-54	V18	SOTSDSOP	Stud. On-Task, Stud. Directed, Stud. Stop, On-Task Probably
	55-59	V19	ATSSD C12	Achievement Test Scores, Stud. Directed
	60-64	V20	ATSTD C13	" " , Tchr. Directed
	65-69	V21	ATSMATH C14	" " , Math
	70-74	V22	ATSREAD C15	" " , Reading
	75-79	V23	ATSTOTAL C16	" " , Total
	80			Blank
CARD 2	1-4	-		Student ID#
	5	-		Blank
	6-7	V24	QUALA C17	Quality of Administration (Items 1-19, 21-31)
	8-9	V25	QUALB	Quality of Administration (Items 8-19, 21-31)
	10-11	-		Blank
	12	V26	FS	# Filmstrips Viewed
	13	V27	PT	# Practice Tests Completed
	14-16	V28	REIN	# Reinforcement Points Earned (_.__)
	17-18	V29	MEANCOPT	Mean Correct on Practice Tests
	19	V30	TITLEI	Students in Title I
	20	V31	SPED	Students in Special Education
	21	V32	ESOL	Students in English as Second Language
	22-24	-		Blank
	25	V33	TSUPP	Teacher Support of Project
	26	V34	TQUALIMP	Teacher Quality of Implementation
	27-28	V35	EVALTOT	Total Score of Project Evaluation
	29-30	V36	EVALFS	Teacher Evaluation of Filmstrips

Columns	Variable	Label	Code Name
31-32	V37	EVALPT	Teacher Evaluation of Practice Tests
33-34	V38	EVALCOMM	Teacher Evaluation of Contact With Staff
35-36	V39	EVALDATA	Teacher Evaluation of Observation and Data Collect
37-38	V40	EVALGEN	Teacher Evaluation of Project in General
39-40	V41	EVALREIN	Teacher Evaluation of Reinforcement (Exp. I Only)
41-42	V42	EVALSPWK	Teacher Evaluation of Spring Workshop (Exp. I Only)
43-44	V43	STWA C18	Student Test Wiseness A
45-46	V44	STWB	Student Test Wiseness B
47-48	V45	STWC	Student Test Wiseness C
49-50	V46	SATT C19	Student Attitude Score

Data From First 29 Cases (2 cards per case) of Title I Data File

```
                                                                            9:29 PM TUESDAY, JUNE 29, 1982
  100   0001              30101 1019220722100096                    -1.27-0.59-0.07-0.83-0.54
  200   0001 5642  9718680100    33181926101714241009060408
  300   0002              30101 1019220722100096                    -0.51 0.35-0.39-0.10-0.00
  400   0002 5642  9727192100    33181926101714241015070608
  500   0003              30101 1019220722100096                     0.43 0.90 1.23 0.97 1.30
  600   0003 5642  9730094000    33181926101714241013070508
  700   0004              30101 1019220722100096                     1.58 0.90 1.58 1.61 1.37
  800   0004 5642  9737196000    33181926101714241017070410
  900   0005              30101 1019220722100096                     0.58 0.09 0.02 0.52-0.12
 1000   0005 5642  9725781000    33181926101714241011060311
 1100   0006              30101 1019220722100096                    -1.54-1.16 0.41-1.38-0.65
 1200   0006 5642  9712958100    33181926101714241004060511
 1300   0007              30101 1019220722100096                     0.14-0.10 0.84-0.07 0.23
 1400   0007 5642  9527590000    33181926101714241012060408
 1500   0008              30101 1019220722100096                     0.14 1.74 1.05 0.81 0.91
 1600   0008 5642  9528090000    33181926101714241012070412
 1700   0009              30101 1019220722100096                     0.95 0.90 1.23 1.25 1.07
 1800   0009 5642  9730090000    33181926101714241018050611
 1900   0010              30101 1019220722100096                    -1.40-0.59-0.71-1.26-1.19
 2000   0010 5642  9714367100    33181926101714241005070608
 2100   0011              30101 1019220722100096                    -0.61-0.49-0.37-0.83-0.70
 2200   0011 5642  9721472000    33181926101714241012060413
 2300   0012            A 30101 1019220722100096                     1.58 1.74 1.05 1.61 1.49
 2400   0012 5642  9738698000    33181926101714241019070610
 2500   0013              30101 1019220722100096                     0.95 0.35 0.84 1.25 1.40
 2600   0013 5642  9728693000    33181926101714241014070612
 2700   0014              30101 1019220722100096                    -0.37 0.35 0.25-0.23 0.09
 2800   0014 5642  9734378000    33181926101714241012060608
 2900   0015              30101 1019220722100096079090054058096100 0.14 0.90 0.01 0.25 0.68
 3000   0015 5642  9741457000    33181926101714241013070510
 3100   0016              30101 1019220722100096061086053064080100-0.72-1.07-0.25-0.89-0.77
 3200   0016 5642  9718676100    33181926101714241006030513
 3300   0017              30101 1019220722100096
 3400   0017 5642  97      000    33181926101714241010060611
 3500   0018              30101 1019220722100096                     0.43-0.25-0.1[illegible] 0.02-0.30
 3600   0018 5642  9728693000    33181926101714241014070509
 3700   0019              30101 1019220722100096086043050091094095-0.51-0.10 0.69-0.42 0.03
 3800   0019 5642  8621780100    33181926101714241011070510
 3900   0020              30101 1019220722100096                    -0.06 0.35 0.01 0.15 0.02
 4000   0020 5642  9630086000    33181926101714241013070508
 4100   0021              30101 1019220722100096                    -1.15-0.69-1.02-1.13-1.30
 4200   0021 5642  9722968100    33181926101714241005060608
 4300   0022              30101 1019220722100096070087080086094100 1.58 1.74 2.05 1.05 1.30
 4400   0022 5642  9738698000    33181926101714241010050508
 4500   0023              30101 1019220722100096                     0.43 0.90 1.23 0.97 1.30
 4600   0023 5642  9731490000    33181926101714241016070508
 4700   0024              30101 1019220722100096076086055055100100 0.95 1.74 2.05 1.25 1.88
 4800   0024 5642  9735796000    33181926101714241019060608
 4900   0025              30101 1019220722100096
 5000   0025 5642  87      000    331819261017142410
 5100   0036              30102 1722240920095078                     1.58 1.74 1.05 1.61 1.23
 5200   0036 5239  9728696000    33191917123019242019020509
 5300   0037              30102 1722240920095078                    -0.06-0.10-0.57-0.06-0.39
 5400   0037 5239  9720077100    33191917123019242006060612
 5500   0038              30102 1722240920095078                     0.14 1.74 0.11 0.81 0.95
 5600   0038 5239  9728691000    33191917123019242014060613
 5700   0039              30102 1722240920095078                     0.14 0.90 1.05-0.14 0.79
 5800   0039 5239  9725791000    33191917123019242009070409
```

BEST COPY AVAILABLE

Table H.1

Means, Medians, and Standard Deviations for Dependent Variables by Experimental Group (All Students)

Variable	GROUP I				GROUP II				CONTROL GROUP				F	P
	N	$\bar{X}$	SD	Median	N	$\bar{X}$	SD	Median	N	$\bar{X}$	SD	Median		
Teacher Attitude (Total)	552	89.1	13.2	89.8	460	87.9	11.1	87.2	474	85.7	12.3	85.5	9.85	.000
Teacher Attitude (Opinion)	552	14.7	4.3	14.8	460	15.1	3.6	15.1	474	14.9	3.1	14.9	1.23	.292
Teacher Attitude (Feeling)	552	20.7	2.6	20.2	460	20.9	2.8	20.7	474	19.5	3.1	19.6	37.32	.000
Teacher Attitude (Use)	552	24.8	4.7	23.3	460	25.6	3.7	26.8	474	25.9	4.2	25.8	9.07	.000
Teacher Attitude (Increase)	552	10.5	2.6	10.9	460	10.0	1.9	10.1	474	10.1	2.5	10.1	7.91	.000
Teacher Attitude (S. Feel)	552	18.3	4.9	18.2	460	16.2	3.9	16.6	474	15.3	3.3	15.5	67.94	.000
Teacher On-Task (TD*)	501	77.1	35.7	98.3	460	73.1	34.7	89.1	432	59.2	49.0	68.9	25.14	.000
Teacher On-Task (SD**)	501	80.5	22.8	92.8	460	78.4	24.3	88.7	432	83.2	37.7	87.7	3.05	.048
Student On-Task (TD)	91	88.4	14.3	90.8	85	89.7	11.8	94.4	90	89.2	11.1	92.5	.24	.785
Student On-Task (SD)	95	90.6	11.9	93.4	79	89.9	11.9	93.8	90	90.5	9.3	92.5	.09	.911
Achievement Test (SD)	512	-.1	1.0	.2	418	.0	1.0	.4	454	.1	.9	.3	3.43	.033
Achievement Test (TD)	512	-.1	1.0	.1	417	.1	1.0	.3	454	.0	.9	.1	3.79	.023
Achievement Test (Math)	513	-.1	1.0	-.1	418	-.1	1.0	-.1	452	.2	.9	.2	11.94	.000
Achievement Test (Total Read)	511	-.1	1.0	.1	417	.0	1.0	.3	454	.1	.9	.2	4.36	.013
Achievement Test (Total)	509	-.1	1.0	.0	417	.0	1.0	.1	452	.1	.9	.3	7.68	.000
Quality of Test Administration	501	49.7	8.5	52.1	460	50.6	3.6	50.9	432	48.8	3.8	48.3	11.42	.000
Test-Taking Skills (Wise)	489	9.9	4.0	9.4	387	10.0	4.0	9.7	462	10.2	4.1	10.0	.88	.412
Test-Taking Skills (Deductive)	489	6.1	1.2	6.3	387	6.1	1.2	6.3	462	6.1	1.3	6.3	.45	.640
Test-Taking Skills (Directions)	489	5.1	1.1	5.5	387	5.2	1.1	5.6	462	5.2	1.1	5.5	1.12	.326
Student Attitude	490	11.9	3.8	11.3	388	11.7	3.4	11.2	468	12.4	3.5	12.1	4.57	.011

*Teacher Directed
**Student Directed

BEST COPY AVAILABLE

Table H.2

Means, Medians, and Standard Deviations for Dependent Variables by Experimental Group (Students Receiving Majority of Treatment[a])

Variable	GROUP I				GROUP II				CONTROL GROUP					
	N	$\overline{X}$	SD	Median	N	$\overline{X}$	SD	Median	N	$\overline{X}$	SD	Median	F	P
Teacher Attitude (Opinion)	372	13.9	3.7	14.3	33[illegible]	15.6	3.8	15.6	406	15.0	3.2	14.9	19.51	.000
Teacher Attitude (Feeling)	372	20.2	2.3	19.9	336	20.7	2.8	20.7	406	19.6	3.1	19.7	15.73	.001
Teacher Attitude (Use)	372	24.3	4.5	23.1	336	25.6	3.8	26.9	406	26.0	4.3	25.9	17.72	.000
Teacher Attitude (Increase)	372	10.3	2.7	10.4	336	10.0	1.7	10.1	406	10.1	2.4	10.1	1.84	.160
Teacher Attitude (S. Feel)	372	17.6	5.0	17.9	336	16.6	4.1	17.1	406	15.3	3.4	15.5	28.84	.000
Teacher On-Task (TD)	330	83.6	26.9	99.0	336	77.0	32.0	95.4	368	59.4	50.9	68.8	37.32	.000
Teacher On-Task (SD)	330	81.6	23.0	93.8	336	81.2	20.2	88.7	368	83.1	40.4	88.0	.41	.665
Student On-Task (TD)	68	89.4	11.1	90.5	64	89.2	12.7	94.8	80	89.5	11.4	94.0	.01	.992
Student On-Task (SD)	68	93.0	6.5	94.5	63	89.5	12.7	93.7	80	91.0	9.7	94.3	2.07	.129
Achievement Test (SD)	358	.1	.9	.4	310	.2	.9	.4	389	.3	.8	.4	3.73	.024
Achievement Test (TD)	358	.1	.9	.3	310	.2	.9	.4	389	.2	.9	.2	.68	.508
Achievement Test (Math)	358	.1	.9	.1	311	.1	10.0	.1	388	.3	.9	.4	8.43	.000
Achievement Test (Total Read)	358	.1	.9	.2	310	.2	.1	.4	389	.3	.8	.4	2.85	.058
Achievement Test (Total)	356	.1	.9	.2	310	.2	.9	.3	388	.3	.8	.4	4.26	.014
Quality of Test Administration	330	52.3	3.4	52.4	336	51.3	3.2	51.6	368	48.8	3.7	48.3	98.82	.000
Test-Taking Skills (Wise)	336	10.5	4.[illegible]	10.1	293	10.3	4.0	10.1	395	10.7	4.0	10.4	.66	.517
Test-Taking Skills (Deductive)	336	6.2	1.[illegible]	6.4	293	6.2	1.1	6.4	395	6.3	1.0	6.4	.16	.856
Test-Taking Skills (Directions)	336	5.3	.9	5.6	293	5.4	1.0	5.7	395	5.3	1.0	5.6	.16	.856
Student Attitude	336	11.5	3.4	11.1	292	11.6	3.3	11.1	401	12.4	3.5	12.2	8.74	.000

[a] Eliminating students who saw less than 5 filmstrips, took less than 3 practice tests, had teachers who were rated low on quality of implementation or support, or were in special education programs, or had English as a second language.

Table H.3

Means, Medians, and Standard Deviations for Dependent Variables by Experimental Group (Students Receiving All of Treatment[b])

Variable	GROUP I				GROUP II				CONTROL GROUP			
	N	$\bar{X}$	SD	Median	N	$\bar{X}$	SD	Median	N	$\bar{X}$	SD	Median
Teacher Attitude (Opinion)	216	13.2	3.8	13.5	166	15.3	3.8	15.4	406	15.0	3.2	14.9
Teacher Attitude (Feeling)	216	20.1	2.0	19.9	166	20.6	2.9	19.8	406	19.6	3.1	19.7
Teacher Attitude (Use)	216	23.1	4.3	21.6	166	25.7	3.7	26.8	406	26.0	4.3	25.9
Teacher Attitude (Increase)	216	10.1	2.6	9.4	166	9.7	1.6	9.7	406	10.1	2.4	10.1
Teacher Attitude (S. Feel)	216	17.2	5.7	16.4	166	15.8	4.1	16.8	406	15.3	3.4	15.5
Teacher On-Task (TD)	191	79.7	28.3	96.0	166	71.5	36.5	95.8	368	59.4	50.9	68.8
Teacher On-Task (SD)	191	79.7	24.8	93.8	166	85.2	20.5	88.9	368	83.1	40.4	88.0
Student On-Task (TD)	47	87.1	11.8	87.3	33	86.8	14.7	91.0	80	89.5	11.4	94.0
Student On-Task (SD)	47	92.3	7.0	93.3	33	87.6	14.5	93.6	80	90.9	9.7	94.3
Achievement Test (SD)	209	.1	.9	.4	157	.3	.7	.5	389	.3	.8	.4
Achievement Test (TD)	209	.1	.9	.2	157	.3	.8	.5	389	.2	.9	.2
Achievement Test (Math)	209	.1	.9	.1	157	.2	1.0	.3	388	.3	.9	.8
Achievement Test (Total Read)	209	.1	.9	.2	157	.4	.8	.5	389	.3	.8	.4
Achievement Test (Total)	206	.1	.9	.2	158	.3	.8	.4	388	.3	.8	.4
Quality of Test Administration	191	52.6	3.6	52.5	166	51.1	3.1	51.6	368	48.8	3.7	48.3
Test-Taking Skills (Wise)	200	10.6	4.2	10.2	151	10.6	3.9	10.3	395	10.7	4.0	10.4
Test-Taking Skills (Deductive)	200	6.2	1.2	6.4	151	6.3	1.1	6.5	395	6.3	1.0	6.4
Test-Taking Skills (Directions)	200	5.3	.9	5.6	151	5.4	1.0	5.7	395	5.3	1.0	5.6
Student Attitude	200	11.6	2.3	11.2	149	11.7	3.2	11.2	401	12.4	3.5	12.2

[b] Eliminating students who saw less than 9 filmstrips, took less than 7 practice tests, had teachers who were rated low on quality of implementation or support, or were in special education programs, or had English as a second language.

BEST COPY AVAILABLE

Table H.4

Means, Medians, and Standard Deviations for Dependent Variables by Experimental Group (Only Title I Students Receiving All of Treatment[b])

Variable	GROUP I				GROUP II				CONTROL GROUP			
	N	$\overline{X}$	SD	Median	N	$\overline{X}$	SD	Median	N	$\overline{X}$	SD	Median
Teacher Attitude (Opinion)	65	15.3	3.6	14.0	48	14.7	3.6	15.0	75	15.5	2.4	15.7
Teacher Attitude (Feeling)	65	20.3	2.1	20.0	48	20.7	2.8	20.2	75	19.4	3.1	19.5
Teacher Attitude (Use)	65	22.9	4.3	21.4	48	24.7	4.5	26.2	75	22.1	4.4	24.6
Teacher Attitude (Increase)	65	10.7	2.5	11.3	48	9.4	1.9	9.7	75	10.0	3.2	10.2
Teacher Attitude (S. Feel)	65	18.1	6.1	19.8	48	16.5	3.1	17.0	75	15.6	2.6	15.9
Teacher On-Task (TD)	51	78.2	30.4	97.1	48	74.5	34.6	96.1	75	55.0	35.3	68.9
Teacher On-Task (SD)	51	78.6	25.9	94.3	48	77.9	21.5	88.0	75	83.4	16.6	87.8
Student On-Task (TD)	14	89.4	8.5	87.5	7	92.1	15.3	98.0	19	88.5	11.9	94.8
Student On-Task (SD)	14	90.8	8.1	91.5	7	93.4	5.8	95.0	19	89.4	7.6	89.3
Achievement Test (SD)	63	-.9	.8	-.8	46	-.1	.9	.2	71	-.4	.9	-.3
Achievement Test (TD)	63	-.7	.7	-.7	46	-.2	.9	-.0	71	-.4	.9	-.3
Achievement Test (Math)	63	-.7	.7	-.6	46	-.4	1.0	-.5	70	-.1	.7	-.2
Achievement Test (Total Read)	63	-.9	.8	-.9	46	-.2	.8	.1	71	-.5	.8	-.4
Achievement Test (Total)	62	-.8	.7	-.7	47	-.3	.8	-.1	70	0.4	.7	-.4
Quality of Test Administration	51	52.1	3.9	52.4	48	50.7	3.6	51.7	75	49.4	4.3	50.3
Test-Taking Skills (Wise)	61	7.7	2.7	7.6	41	8.3	3.0	8.4	72	7.7	2.8	7.3
Test-Taking Skills (Deductive)	61	5.8	1.4	6.0	41	6.0	1.6	6.3	72	5.8	1.2	6.0
Test-Taking Skills (Directions)	61	5.1	1.1	5.5	41	5.3	1.0	5.6	72	4.8	1.4	5.1
Student Attitude	61	11.2	3.0	11.0	40	11.5	3.2	11.0	73	12.3	3.8	11.7

[b] Eliminating students who saw less than 9 filmstrips, took less than 7 practice tests, had teachers who were rated low on quality of implementation or support, or were in special education programs, or had English as a second language.

BEST COPY AVAILABLE

Table H.6

Means, Medians, and Standard Deviations for Dependent Variables by Experimental Group (Only Cache and Nebo Districts with Students Receiving Majority of Treatment[a])

Variable	GROUP I				GROUP II				CONTROL GROUP			
	N	$\bar{X}$	SD	Median	N	$\bar{X}$	SD	Median	N	$\bar{X}$	SD	Median
Teacher Attitude (Opinion)	133	13.5	5.2	14.4	121	16.2	2.2	16.4	254	15.1	3.9	14.7
Teacher Attitude (Feeling)	133	19.8	2.7	19.0	121	19.5	1.8	19.0	254	19.2	3.2	18.7
Teacher Attitude (Use)	133	24.6	4.5	23.8	121	27.1	2.8	28.2	254	25.9	4.3	26.2
Teacher Attitude (Increase)	133	9.5	2.2	9.0	121	10.8	1.2	11.2	254	10.3	1.7	10.1
Teacher Attitude (S. Feel)	133	16.7	5.7	18.7	121	15.8	5.7	16.0	254	15.2	3.9	15.1
Teacher On-Task (TD)	133	90.2	16.4	98.3	121	93.6	5.5	96.0	216	71.9	54.9	77.6
Teacher On-Task (SD)	133	75.7	26.6	79.0	121	90.9	17.0	98.0	216	83.2	51.8	93.5
Student On-Task (TD)	28	89.1	9.6	87.5	22	90.3	12.1	95.5	44	89.7	10.9	92.5
Student On-Task (SD)	28	96.6	4.0	97.9	22	85.4	19.5	98.5	44	92.1	11.0	97.0
Achievement Test (SD)	128	.2	.9	.1	116	-.2	1.0	.0	248	.3	.8	.3
Achievement Test (TD)	128	.2	.9	.1	116	-.1	1.0	.1	248	.2	.9	.1
Achievement Test (Math)	128	.1	.9	.1	117	-.2	.9	-.1	248	.3	.9	.4
Achievement Test (Total Read)	128	.2	.9	.1	116	-.1	1.0	.1	248	.3	.8	.4
Achievement Test (Total)	127	.2	.9	.2	116	-.1	1.0	-.0	248	.3	.8	.3
Quality of Test Administration	133	52.9	1.8	52.4	121	51.7	2.8	51.0	216	48.2	3.4	48.4
Test-Taking Skills (Wise)	121	10.8	4.0	10.7	105	10.0	4.2	9.7	247	11.0	4.1	11.2
Test-Taking Skills (Deductive)	121	6.4	1.0	6.6	105	6.2	1.0	6.3	247	6.4	.9	6.6
Test-Taking Skills (Directions)	121	5.2	1.0	5.5	105	5.5	.9	5.7	247	5.4	.9	5.6
Student Attitude	121	11.6	3.5	11.4	105	11.5	3.0	11.2	252	12.6	3.5	12.5

[a] Eliminating students who saw less than 5 filmstrips, took less than 3 practice tests, had teachers who were rated low on quality of implementation or support, or were in special education programs, or had English as a second language.

BEST COPY AVAILABLE

Table H.7

FREQUENCIES ON ALL VARIABLES

Teacher Attitude (Opinion)	$\bar{X}$ = 14.91 SD = 3.74 Mdn = 14.98 Min = 6.00 Max = 25.00	Achievement Test (Math)	$\bar{X}$ = 0.00 SD = .99 Mdn = .02 Min = -3.05 Max = 2.64
Teacher Attitude (Feeling)	$\bar{X}$ = 20.36 SD = 2.08 Mdn = 20.15 Min = 15.00 Max = 25.00	Achievement Test (Total Reading)	$\bar{X}$ = 0 SD = .99 Mdn = .21 Min = -3.48 Max = 1.64
Teacher Attitude (Use)	$\bar{X}$ = 25.43 SD = 4.26 Mdn = 25.48 Min = 16.00 Max = 34.00	Achievement Test (Total)	$\bar{X}$ = 0 SD = .99 Mdn = .14 Min = -3.94 Max = 2.02
Teacher Attitude (Increased Use of Tests)	$\bar{X}$ = 10.23 SD = 2.36 Mdn = 10.26 Min = 4.00 Max = 15.00	Quality of Test Administration	$\bar{X}$ = 49.69 SD = 5.94 Mdn = 50.80 Min = 18.00 Max = 58.00
Teacher Attitude (Students' Feeling)	$\bar{X}$ = 16.70 SD = 4.33 Mdn = 16.57 Min = 6.00 Max = 28.00	Student Test-Taking Skills (Wise)	$\bar{X}$ = 10.02 SD = 4.04 Mdn = 9.67 Min = 0 Max = 23
Teacher On-Task (Teacher-Directed Test)	$\bar{X}$ = 70.24 SD = 40.69 Mdn = 89.41 Min = 0.00 Max = 78.00	Student Test-Taking Skills (Deductive)	$\bar{X}$ = 6.08 SD = 1.23 Mdn = 6.28 Min = 0 Max = 8
Teacher On-Task (Student-Directed Test)	$\bar{X}$ = 80.68 SD = 28.72 Mdn = 89.83 Min = 14.00 Max = 74.00	Student Test-Taking Skills Directions	$\bar{X}$ = 5.18 SD = 1.12 Mdn = 5.54 Min = 0 Max = 7
Student On-Task (Teacher-Directed Test)	$\bar{X}$ = 89.06 SD = 12.47 Mdn = 93.00 Min = 0.00 Max = 100.00	Student Attitude	$\bar{X}$ = 12.03 SD = 3.59 Mdn = 11.59 Min = 0 Max = 24
Student On-Task (Student-Directed Test)	$\bar{X}$ = 90.35 SD = 11.03 Mdn = 93.55 Min = 14.0 Max = 100.0	Student in Title I?	No 0 = 1073 72% Yes 1 = 408 28%
Achievement Test Score (Student-Directed)	$\bar{X}$ = 0.00 SD = .99 Mdn = .29 Min = -3.65 Max = 1.58	Student in Special Education?	No 0 = 1344 91% Yes 1 = 137 9%
Achievement Test Score (Teacher-Directed)	$\bar{X}$ = .00 SD = .99 Mdn = .14 Min = -4.25 Max = 1.99	Student With English as a Second Language?	No 0 = 1439 97% Yes 1 = 42 3%

BEST COPY AVAILABLE

Table H.7 (cont'd)

Teacher Support of Program	1 = 51 5% 2 = 377 37% 3 = 584 58%	$\bar{X}$ = 2.53 SD = .59 Mdn = 2.63 Min = 1 Max = 3	Evaluation of General Impressions	$\bar{X}$ = 2.01 SD = .69 Mdn = 1.99 Min = 1 Max = 4.2
Quality of Teacher Implementation	Poor 1 = 160 16% ↕ 2 = 270 27% Good 3 = 582 58%	$\bar{X}$ = 2.42 SD = .75 Mdn = 2.63 Min = 1 Max = 3	Evaluation of Reinforcement	$\bar{X}$ = 2.83 SD = 1.04 Mdn = 2.97 Min = 1.0 Max = 4.4
Evaluation of Project		$\bar{X}$ = 1.97 SD = .55 Mdn = 1.97 Min = 1 Max = 3.4	Evaluation of Spring Workshop	$\bar{X}$ = 2.18 SD = 1.30 Mdn = 1.88 Min = 1 Max = 5
Evaluation of Filmstrips		$\bar{X}$ = 1.65 SD = .49 Mdn = 1.62 Min = 1 Max = 3	No. of Filmstrips Viewed	$\bar{X}$ = 8.19 SD = 1.46 Mdn = 8.67 Min = 0 Max = 9
Evaluation of Practice Tests		$\bar{X}$ = 2.04 SD = .75 Mdn = 1.98 Min = 1 Max = 4.4	No. of Practice Tests Taken	$\bar{X}$ = 6.34 SD = 1.14 Mdn = 6.68 Min = 0 Max = 7
Evaluation of Communication		$\bar{X}$ = 1.65 SD = .62 Mdn = 1.72 Min = .80 Max = 3.0	$\bar{X}$ No. of Reinforcement Points Per Test	$\bar{X}$ = 3.77 SD = 1.44 Mdn = 3.67 Min = .40 Max = 9.5
Evaluation of Data Collection		$\bar{X}$ = 2.07 SD = .89 Mdn = 2.06 Min = 1 Max = 4.7	Mean % Correct on Practice Tests	$\bar{X}$ = 82.75% SD = 14.72% Mdn = 87.18% Min = 7% Max = 99%

BEST COPY AVAILABLE